Mechanism Design for LLM Fine-tuning with Multiple Reward Models

Haoran Sun
Peking University
sunhaoran0301@stu.pku.edu.cn

Yurong Chen
Peking University
chenyurong@pku.edu.cn

Siwei Wang
Microsoft Research Asia
siweiwang@microsoft.com

Wei Chen
Microsoft Research Asia
weic@microsoft.com

Xiaotie Deng
Peking University
xiaotie@pku.edu.cn

Abstract

Recent research on fine-tuning large language models (LLMs) through the aggregation of multiple preferences has attracted considerable attention. However, the existing literature predominantly focuses on the empirical performance of aggregation algorithms while neglecting the underlying motivation for agents to misreport their preferences. In this paper, we formalize this as a multi-parameter mechanism design problem, where an LLM provider designs training and payment rules to achieve specific objectives and promote the truthful reporting of preferences. Firstly, we claim the necessity of a payment scheme by demonstrating that without payments, truth-telling is a strictly dominated strategy under a wide range of training rules. Then, we introduce the affine maximizer payment scheme for the social welfare maximizing training rules, which ensures both dominant-strategy incentive compatibility (DSIC) and individual rationality (IR). Furthermore, we prove that under mild conditions, any other payment rule that implements these training rules in DSIC can be converted to the affine maximizer payment by adding a factor irrelevant to the agents' reports. We also show that this mechanism satisfies approximate DSIC when the input of the mechanism is a biased version of the reported preferences, showcasing its robustness in real-world applications.

1 Introduction

The process of fine-tuning an LLM to align with specific human preferences is challenging to achieve through supervision (Ji et al. [2023], Köpf et al. [2024], Wang et al. [2023b], Shen et al. [2023]), primarily due to the difficulty in constructing datasets with a substantial number of valid question-answer pairs for supervised training. Reinforcement learning from human feedback (RLHF) (Ouyang et al. [2022], Christiano et al. [2017]) offers a promising solution to this problem. In RLHF, a reward model is first trained as a proxy for human judgment. This model then provides reward signals for the standard reinforcement learning process. This fine-tuning technique with a reward model has proven effective in encoding human preferences into models and has become a fundamental component of the training process for most advanced LLMs. With the advancement of RLHF, numerous studies have investigated efficient methods for aggregating multiple preferences into a single fine-tuned model.

However, most of these studies focus on improving empirical performance across various metrics (Ramé et al. [2024], Wu et al. [2024], Coste et al. [2023], Zhang et al. [2024a], Jang et al. [2023], Eisenstein et al. [2023], Yang et al. [2024], Rame et al. [2024], Shi et al. [2024]). They often implicitly assume that we are accessible to actual preferences, neglecting the possibility of agents' misreporting their preferences. This problem becomes more crucial when considering a real-world scenario where

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

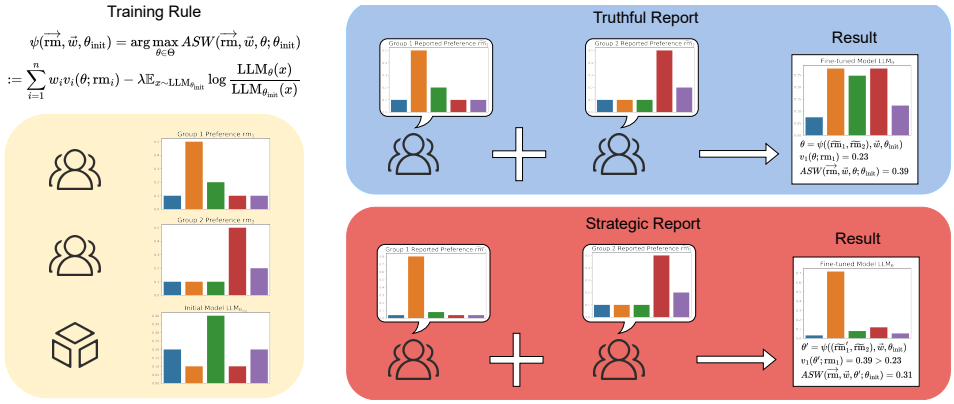

Figure 1: Motivating example of the RLHF Game: Consider a basic training rule ψ in RLHF for two groups, setting $w_1 = w_2 = 1$. When there is no payment rule and group 2's report $\widetilde{\mathrm{rm}}_2$ is fixed, the valuation that group 1 can achieve from a truthful report $\widetilde{\mathrm{rm}}_1 = \mathrm{rm}_1$, $v_1(\theta; \mathrm{rm}_1)$, is strictly less than from a strategic report $\widetilde{\mathrm{rm}}'_1 \neq \mathrm{rm}_1$, $v_1(\theta'; \mathrm{rm}_1)$. On the other hand, such strategic behavior also has an impact on the affine social welfare: $ASW(\overrightarrow{\mathrm{rm}}, \vec{w}, \theta; \theta_{\text{init}}) < ASW(\overrightarrow{\mathrm{rm}}, \vec{w}, \theta'; \theta_{\text{init}})$.

different agents provide their preferences for the aggregation. In such cases, agents may engage in strategic misreporting to increase utility. An intuitive example is if an agent knows beforehand that the fine-tuning process aims to neutralize all preferences, it might pretend to have a more polarized preference as a beneficial strategy, as shown in Figure 1. These strategic behaviors can distort the final training results, even if the trained algorithm is highly effective. Nevertheless, this issue has not attracted sufficient attention in the existing literature, particularly concerning the fine-tuning process of LLMs.

Our Contribution. In this paper, we mainly study the incentive design in such scenarios. First, we formalize this as a multi-parameter mechanism design problem, which we call the *RLHF Game*, involving a fine-tuning service provider and groups of agents seeking the service.

Next, we demonstrate the necessity of payment mechanisms for commonly used SW-Max training rules (Theorem 3.2) and derive that the affine maximizer payment scheme can implement these rules in both dominant-strategy incentive compatibility (DSIC) and individual rationality (IR) (Theorem 3.3).

We further explore payment equivalence, showing that under a mild condition, any other payment rule that also implements these training rules in DSIC can be converted to the affine maximizer payment by adding a factor irrelevant to groups' reports (Theorem 3.5). Consequently, we derive the revenue-maximizing payment rule that implements SW-Max training rules in both DSIC and IR (Corollary 3.6).

Finally, we show that the mechanism remains approximately DSIC even when the input preferences are biased, reflecting practical scenarios where errors occur (Theorem 3.7). We also provide preliminary empirical validation in real RLHF scenarios Appendix B.2.

2 Formulation of the RLHF Game

In this section, we present the formal description of the RLHF Game. In the RLHF Game, there is one LLM *provider* and n *groups of agents*, denoted by $[n] = \{1, 2, \cdots, n\}$. Let $T^* := \emptyset \cup T \cup T^2 \cup \cdots \cup T^K$ represent the set of all possible input sequences with lengths up to K. The provider has an initial model $\mathrm{LLM}_{\theta_{\text{init}}}$ with non-zero probability for all sequences, i.e., $\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in T^*$. We mainly consider two types of reward models: normalized by summation ($\sum_{x \in T^*} \mathrm{rm}(x) = 1$) and normalized by maximum ($\max_{x \in T^*} \mathrm{rm}(x) = 1$). Each group i has w_i agents and a joint preference represented by a reward model $\mathrm{rm}_i : T^* \to \mathbb{R}_{\geq 0}$. Let $\mathcal{R}$ and $\mathcal{W} \subseteq \mathbb{N}_+$ denote the domains for each group's reward model and group size, respectively. We assume an upper bound $\bar{w}$ for $\mathcal{W}$. The exact reward model and the size are group i's private information. For an agent in group i, the valuation when it receives a model LLM_θ is denoted by $v_i(\theta; \mathrm{rm}_i)$. We consider a reasonable form $v(\cdot; \cdot)$:

Definition 2.1. For any agent with preference represented by reward model rm, its valuation on model LLM_θ is its expected reward on the sequences generated by LLM_θ: $v(\theta; \text{rm}) = \mathbb{E}_{\boldsymbol{x}\sim\text{LLM}_\theta}\text{rm}(\boldsymbol{x}) = \sum_{\boldsymbol{x}\in T^*}\text{LLM}_\theta(\boldsymbol{x})\text{rm}(\boldsymbol{x})$.

The provider first announces the mechanism, including a training rule $\psi : \mathcal{R}^n \times \mathcal{W}^n \times \Theta \to \Theta$ and a payment rule $p : \mathcal{R}^n \times \mathcal{W}^n \times \Theta \to \mathbb{R}^n$. Both rules take n reported reward models, n reported sizes, and an initial model as input and output the objective fine-tuned model and each group's payment, respectively. Specifically, the training rule seeks to find the model that maximizes a specific objective function h. That is, $\psi(\overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}) \in \arg\max_{\theta\in\Theta} h(v_1(\theta; \text{rm}_1), \cdots, v_n(\theta; \text{rm}_n), \vec{w}, D(\text{LLM}_\theta || \text{LLM}_{\theta_{\text{init}}}))$, where D is a measure of the distance between LLM_θ and $\text{LLM}_{\theta_{\text{init}}}$. We assume the function h has a unique global optimal point for any possible inputs. Hence, in the definition of ψ, we use "=" to substitute "$\in$".

After observing the announced mechanism (ψ, p), each group i reports a reward model, $\widetilde{\text{rm}}_i$, and its group size $\tilde{w}_i \le \bar{w}$. Based on the reported information, the provider fine-tunes the model and gets the final parameter $\theta_{\text{final}} = \psi(\overrightarrow{\widetilde{\text{rm}}}, \vec{\tilde{w}}, \theta_{\text{init}})$. We assume each group i has a quasi-linear utility, which means $u_i(\overrightarrow{\widetilde{\text{rm}}}, \vec{\tilde{w}}; \psi, p, \text{rm}_i, w_i) = w_i v_i(\theta_{\text{final}}; \text{rm}_i) - p_i(\overrightarrow{\widetilde{\text{rm}}}, \vec{\tilde{w}}, \theta_{\text{init}})$. When the specific mechanism (ψ, p) is given, we will omit their notations for simplicity.

The goal of the LLM provider is to achieve its training objective based on the group's true preferences, taking into account that the misreporting may distort the training outcome. To this end, it is crucial to incentivize all groups to report their information truthfully so that the provider is accessible to the groups' private information. We formally define these desiderata of a mechanism:

(1) A mechanism (ψ, p) satisfies $\epsilon-$*dominant-strategy incentive compatibility* (ϵ-DSIC) if $\forall i$, rm_i, w_i, rm'_i, w'_i, $\overrightarrow{\text{rm}}_{-i}$, $\vec{w}_{-i}$, θ_{init}, we have

$$u_i((\text{rm}_i, \overrightarrow{\text{rm}}_{-i}), (w_i, \vec{w}_{-i}); \text{rm}_i, w_i) + \epsilon \ge u_i((\text{rm}'_i, \overrightarrow{\text{rm}}_{-i}), (w'_i, \vec{w}_{-i}); \text{rm}_i, w_i). \qquad (\epsilon\text{-DSIC})$$

(2) A mechanism (ψ, p) satisfies $\epsilon-$*individually rationality* (ϵ-IR) if $\forall i$, rm_i, w_i, $\overrightarrow{\text{rm}}_{-i}$, $\vec{w}_{-i}$, θ_{init}, we have

$$u_i((\text{rm}_i, \overrightarrow{\text{rm}}_{-i}), (w_i, \vec{w}_{-i}); \text{rm}_i, w_i) + \epsilon \ge 0. \qquad (\epsilon\text{-IR})$$

In particular, we use the terms DSIC and IR to refer to 0-DSIC and 0-IR, respectively. When a mechanism (ψ, p) satisfies DSIC, IR, or both DSIC and IR, we say that the payment rule p *implements* ψ in DSIC, IR or both DSIC and IR. When we say the implementability of a training rule, we refer to the property of DSIC.

3 Incentives for SW-Maximizing Training Rules

This section will discuss the incentive design within the RLHF Game framework. Our primary focus is on a subset of training rules that maximizes social welfare under regularization constraints, which is commonly used in practice to aggregate various preferences (Boyd and Vandenberghe [2004], Nocedal and Wright [1999]).

Definition 3.1 (SW-Max Training Rules)**.** A SW-Max training rule fine-tunes the model to maximize social welfare, subject to a regularization penalty measured by f-divergence (Ali and Silvey [1966], Csiszár [1967], Shi et al. [2024]). Formally, this can be expressed as: $\psi(\overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}) = \arg\max_{\theta\in\Theta} \sum_{i=1}^n w_i v_i(\theta; \text{rm}_i) - \lambda \mathbb{E}_{\boldsymbol{x}\sim\text{LLM}_{\theta_{\text{init}}}} f(\text{LLM}_\theta(\boldsymbol{x})/\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}))$, where f is convex on $\mathbb{R}_+$ and $f(1) = 0$, and $\lambda > 0$ is the hyperparameter that controls regularization strength. We use $\psi \in \Psi^{SW}$ to indicate that ψ is a member of the SW-Max training rules.

3.1 Necessity of Payment Rule

We begin by showing that without payment, strategies always exist that bring strictly higher utility than truthful reports for SW-Max training rules.

Theorem 3.2. *For the mechanism (ψ, p) that $\psi \in \Psi^{SW}$ and $p \equiv 0$, assuming that for all $\vec{w}$, $\overrightarrow{rm}$ and θ_{init}, the fine-tuned model $\theta = \psi(\overrightarrow{rm}, \vec{w}, \theta_{init})$ satisfies that $LLM_\theta(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in T^*$, then for group i, truthfully reporting is a* strongly dominated strategy *when $\min_{\boldsymbol{x}\in T^*} rm_i(\boldsymbol{x}) > 0$ and $|\{r | r = rm_i(x), x \in T^*\}| \ge 2$.*

Here, we call a strategy strongly dominated when there exists another strategy that yields *strictly higher utility* regardless of others' reports. Theorem 3.2 tells us that truthful reporting is strongly dominated with only training rules and thus will not be adopted by rational groups. We prove this result by constructing such report reward models rm_i', and the intuitive is that rm_i' assigns a lower value for the less preferred $\boldsymbol{x}$ and a higher value for the most preferred $\boldsymbol{x}$.

3.2 Affine Maximizer Payment

Having established the necessity of payment rules in this scenario, we mainly address two questions in this part: First, *given a training rule ψ, can we find a payment rule p such that the mechanism (ψ, p) satisfies DSIC?* This is the so-called implementability of a training rule ψ. Second, *for an implementable training rule ψ, can we identify the relationship between the payment rules ps among all DSIC mechanisms (ψ, p).*

For SW-Max training rules, we resolve the first question by introducing affine maximizer payment rule (Roberts [1979]), which is a weighted version of the well-known VCG payment (Vickrey [1961], Clarke [1971], Groves [1973]). The payment rule can be written as $p_i^{AFF}(\overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}) = ASW_{-i}(\overrightarrow{\text{rm}}, \vec{w}, \psi(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}}); \theta_{\text{init}}) - ASW_{-i}(\overrightarrow{\text{rm}}, \vec{w}, \psi(\overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}); \theta_{\text{init}})$, where the notation $ASW_{-j}(\overrightarrow{\text{rm}}, \vec{w}, \theta; \theta_{\text{init}}) := \sum_{i \neq j} w_i v_i(\theta; \text{rm}_i) - \lambda \mathbb{E}_{\boldsymbol{x} \sim \text{LLM}_{\theta_{\text{init}}}} f(\text{LLM}_\theta(\boldsymbol{x}) / \text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}))$ refer to the affine social welfare without group j for a specific fine-tuned model. We show that p^{AFF} implements SW-Max training rules in both DSIC and IR, implying that truthfully reporting both reward models and group sizes constitutes a dominant Nash Equilibrium in this mechanism.

Theorem 3.3. *For any $\psi \in \Psi^{SW}$, mechanism (ψ, p^{AFF}) satisfies DSIC and IR.*

The second question is more general, so we primarily consider the concept of *payment equivalence* ([Ashlagi et al., 2010]) defined as:

Definition 3.4 (Payment Equivalence)**.** An implementable training rule ψ satisfies payment equivalence if for any two mechanisms (ψ, p) and (ψ, p') satisfying DSIC, there exists a function f such that for $\forall \text{rm}_i \in \mathcal{R}_i$, $p_i'(\text{rm}_i, \overrightarrow{\text{rm}}_{-i}; \theta_{\text{init}}) = p_i(\text{rm}_i, \overrightarrow{\text{rm}}_{-i}; \theta_{\text{init}}) + f(\overrightarrow{\text{rm}}_{-i}, \theta_{\text{init}})$.

Payment equivalence indicates that the only way to modify a DSIC mechanism (ψ, p) to (ψ, p') while maintaining incentive compatibility is to add a term that is independent of i's report to group i's payment function p_i. Thus, the payment equivalence of ψ is sometimes interpreted as the uniqueness of the payment rule p that implements it in DSIC. This notion is strong and useful since when a training rule ψ satisfies payment equivalence, and we can figure out one mechanism (ψ, p) that satisfies DSIC, all the payment rules p' that implement ψ in DSIC are characterized. We show that SW-Max training rules satisfy such property under a mild assumption, which holds for various distance measures (see Proposition C.2).

Theorem 3.5. *When for any $\epsilon > 0$, there exists a $\delta > 0$ such that for any θ_{init}, $\overrightarrow{rm}$, $\overrightarrow{rm}'$, $\vec{w}$ and $\vec{w}'$, if $\max_{\boldsymbol{x} \in T^*} |\sum_{i=1}^n (w_i rm_i(\boldsymbol{x}) - w_i' rm_i'(\boldsymbol{x}))| \leq \delta$, then $\max_{\boldsymbol{x} \in T^*} |LLM_\theta(\boldsymbol{x}) - LLM_{\theta'}(\boldsymbol{x})| \leq \epsilon$, where $\theta := \psi(\overrightarrow{rm}, \vec{w}, \theta_{init})$ and $\theta' := (\overrightarrow{rm}', \vec{w}', \theta_{init})$, each training rule $\psi \in \Psi^{SW}$ satisfies payment equivalence.*

With the property of payment equivalence, we can investigate the revenue-maximizing payment rule that implements SW-Max training rules in both DSIC and IR.

Corollary 3.6. *Under the assumption in Theorem 3.5, for each training rule $\psi \in \Psi^{SW}$, the revenue-maximizing payment rule p^* under a distribution F whose support is $\mathcal{R} \times \mathcal{W}$ that implements ψ in both DSIC and IR is given by*

$$p_i^*(\overrightarrow{rm}, \vec{w}, \theta_{init}) = p_i^{AFF}(\overrightarrow{rm}, \vec{w}, \theta_{init}) + \inf_{rm_i' \in \mathcal{R}, w_i' \in \mathcal{W}} u_i((rm_i', \overrightarrow{rm}_{-i}), (w_i', \vec{w}_{-i}); \psi, p^{AFF}, rm_i', w_i').$$

Finally, we discuss the influence of error generated in practice on the incentive property in the RLHF Game. We abstract it as an approximate valuation problem (Chiesa et al. [2012]). Formally, when group i reports its reward model rm_i, the mechanism will take a noisy reward model $\widehat{\text{rm}}_i$ with a conditional distribution $F_i(\cdot|\text{rm}_i)$ as the input into the mechanism. For simplicity, we assume that each group only considers such randomness for itself. Thus, under mechanism (ψ, p), the expected utility of group i is given by $U_i((\text{rm}_i', \overrightarrow{\text{rm}}_{-i}), (w_i', \vec{w}_{-i}); \text{rm}_i, w_i) = \mathbb{E}_{\widehat{\text{rm}}_i \sim F_i(\cdot|\text{rm}_i')} u_i((\widehat{\text{rm}}_i, \overrightarrow{\text{rm}}_{-i}), (w_i', \vec{w}_{-i}); \text{rm}_i, w_i)$. We derive the following connection between the magnitude of the error and the deviation from DSIC.

Theorem 3.7. *In the approximate valuation model, assuming* $\max_{\boldsymbol{x} \in T^*, \widehat{rm}_i \sim F_i(\cdot|rm_i)} |\widehat{rm}_i(\boldsymbol{x}) - rm_i(\boldsymbol{x})| \leq \epsilon$ *for all* $i \in [n]$*, when* $\vec{w}$ *is truthfully reported, the mechanism* (ψ, p^{AFF}) *that* $\psi \in \Psi^{SW}$ *is* $\max_{i \in [n]} 2w_i\epsilon$*-DSIC.*

This theorem means that for any group i, the maximum gain of misreporting is less than $2w_i\epsilon$ regardless of the others' reports. Groups will tend to truthfully report in cases where finding the optimal strategy is costlier than $2w_i\epsilon$.

References

Saaket Agashe, Yue Fan, and Xin Eric Wang. Evaluating multi-agent coordination abilities in large language models, 2023.

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models, 2023.

Syed Mumtaz Ali and Samuel D Silvey. A general class of coefficients of divergence of one distribution from another, 1966.

Itai Ashlagi, Mark Braverman, Avinatan Hassidim, and Dov Monderer. Monotonicity and implementability, 2010.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.

Dirk Bergemann and Juuso Välimäki. The dynamic pivot mechanism, 2010.

Sushil Bikhchandani, Shurojit Chatterji, Ron Lavi, Ahuva Mu'alem, Noam Nisan, and Arunava Sen. Weak monotonicity characterizes deterministic dominant-strategy implementation, 2006.

Stephen P Boyd and Lieven Vandenberghe. Convex optimization, 2004.

Patrick Briest, Shuchi Chawla, Robert Kleinberg, and S Matthew Weinberg. Pricing randomized allocations, 2010.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. Maxmin-rlhf: Towards equitable alignment of large language models with diverse human preferences, 2024.

Yiting Chen, Tracy Xiao Liu, You Shan, and Songfa Zhong. The emergence of economic rationality of gpt, 2023.

Alessandro Chiesa, Silvio Micali, and Zeyuan Allen Zhu. Mechanism design with approximate valuations, 2012.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2017.

Edward H Clarke. Multipart pricing of public goods, 1971.

Vincent Conitzer and Tuomas Sandholm. Self-interested automated mechanism design and implications for optimal combinatorial auctions, 2004.

Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H Holliday, Bob M Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, et al. Social choice for ai alignment: Dealing with diverse human feedback, 2024.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization, 2023.

Imre Csiszár. On information-type measure of difference of probability distributions and indirect observations, 1967.

Michael Curry, Tuomas Sandholm, and John Dickerson. Differentiable economics for randomized affine maximizer auctions, 2022.

Zhijian Duan, Haoran Sun, Yurong Chen, and Xiaotie Deng. A scalable neural network for dsic affine maximizer auction design, 2024a.

Zhijian Duan, Haoran Sun, Yichong Xia, Siqiang Wang, Zhilin Zhang, Chuan Yu, Jian Xu, Bo Zheng, and Xiaotie Deng. Scalable virtual valuations combinatorial auction design by combining zeroth-order and first-order optimization method, 2024b.

Kumar Avinava Dubey, Zhe Feng, Rahul Kidambi, Aranyak Mehta, and Di Wang. Auctions with llm summaries, 2024.

Paul Duetting, Vahab Mirrokni, Renato Paes Leme, Haifeng Xu, and Song Zuo. Mechanism design for large language models, 2023.

Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alex D'Amour, DJ Dvijotham, Adam Fisch, Katherine Heller, Stephen Pfohl, Deepak Ramachandran, et al. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking, 2023.

Meta Fundamental AI Research Diplomacy Team (FAIR)†, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning, 2022.

Caoyun Fan, Jindou Chen, Yaohui Jin, and Hao He. Can large language models serve as rational players in game theory? a systematic analysis, 2023.

Soheil Feizi, MohammadTaghi Hajiaghayi, Keivan Rezaei, and Suho Shin. Online advertisements with llms: Opportunities and challenges, 2023.

Xidong Feng, Yicheng Luo, Ziyan Wang, Hongrui Tang, Mengyue Yang, Kun Shao, David Mguni, Yali Du, and Jun Wang. Chessgpt: Bridging policy learning and language modeling, 2024.

Roberto Gallotta, Graham Todd, Marvin Zammit, Sam Earle, Antonios Liapis, Julian Togelius, and Georgios N Yannakakis. Large language models and games: A survey and roadmap, 2024.

Kanishk Gandhi, Dorsa Sadigh, and Noah D Goodman. Strategic reasoning with language models, 2023.

Ian Gemp, Yoram Bachrach, Marc Lanctot, Roma Patel, Vibhavari Dasagi, Luke Marris, Georgios Piliouras, and Karl Tuyls. States as strings as strategies: Steering language models with game-theoretic solvers, 2024.

Theodore Groves. Incentives in teams, 1973.

Shangmin Guo, Haochuan Wang, Haoran Bu, Yi Ren, Dianbo Sui, Yu-Ming Shang, and Siting Lu. Large language models as rational players in competitive economics games, 2023.

Shangmin Guo, Haoran Bu, Haochuan Wang, Yi Ren, Dianbo Sui, Yuming Shang, and Siting Lu. Economics arena for large language models, 2024a.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges, 2024b.

Birgit Heydenreich, Rudolf Müller, Marc Uetz, and Rakesh V Vohra. Characterization of revenue equivalence, 2009.

Radosveta Ivanova-Stenzel and Timothy C Salmon. Revenue equivalence revisited, 2008.

Athul Paul Jacob, Yikang Shen, Gabriele Farina, and Jacob Andreas. The consensus game: Language model generation via equilibrium search, 2023.

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. Personalized soups: Personalized large language model alignment via post-hoc parameter merging, 2023.

Philippe Jehiel, Moritz Meyer-Ter-Vehn, and Benny Moldovanu. Mixed bundling auctions, 2007.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. Ai alignment: A comprehensive survey, 2023.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations-democratizing large language model alignment, 2024.

Benjamin Laufer, Jon Kleinberg, and Hoda Heidari. Fine-tuning games: Bargaining and adaptation for general-purpose models, 2023.

Anton Likhodedov and Tuomas Sandholm. Methods for boosting revenue in combinatorial auctions, 2004.

Nunzio Lorè and Babak Heydari. Strategic behavior of large language models: Game structure vs. contextual framing, 2023.

David G Luenberger, Yinyu Ye, et al. Linear and nonlinear programming, 1984.

Weiyu Ma, Qirui Mi, Xue Yan, Yuqiao Wu, Runji Lin, Haifeng Zhang, and Jun Wang. Large language models play starcraft ii: Benchmarks and a chain of summarization approach, 2023.

Mitsunobu Miyake. On the incentive properties of multi-item auctions, 1998.

Gabriel Mukobi, Hannah Erlebach, Niklas Lauffer, Lewis Hammond, Alan Chan, and Jesse Clifton. Welfare diplomacy: Benchmarking language model cooperation, 2023.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. Nash learning from human feedback, 2023.

Roger B Myerson. Optimal auction design, 1981.

Jorge Nocedal and Stephen J Wright. Numerical optimization, 1999.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback, 2022.

Chanwoo Park, Mingyang Liu, Kaiqing Zhang, and Asuman Ozdaglar. Principled rlhf from heterogeneous feedback via personalization and preference aggregation, 2024.

Alessandro Pavan, Ilya Segal, and Juuso Toikka. Dynamic mechanism design: A myersonian approach, 2014.

Alexandre Rame, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards, 2024.

Alexandre Ramé, Nino Vieillard, Léonard Hussenot, Robert Dadashi, Geoffrey Cideron, Olivier Bachem, and Johan Ferret. Warm: On the benefits of weight averaged reward models, 2024.

Kevin Roberts. The characterization of implementable choice rules, 1979.

Jean-Charles Rochet. A necessary and sufficient condition for rationalizability in a quasi-linear context, 1987.

Corby Rosset, Ching-An Cheng, Arindam Mitra, Michael Santacroce, Ahmed Awadallah, and Tengyang Xie. Direct nash optimization: Teaching language models to self-improve with general preferences, 2024.

Michael Saks and Lan Yu. Weak monotonicity suffices for truthfulness on convex domains, 2005.

Tuomas Sandholm and Anton Likhodedov. Automated design of revenue-maximizing combinatorial auctions, 2015.

Xiao Shao, Weifu Jiang, Fei Zuo, and Mengqing Liu. Swarmbrain: Embodied agent for real-time strategy game starcraft ii via large language models, 2024.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. Large language model alignment: A survey, 2023.

Ruizhe Shi, Yifang Chen, Yushi Hu, ALisa Liu, Noah Smith, Hannaneh Hajishirzi, and Simon Du. Decoding-time language model alignment with multiple objectives, 2024.

Ermis Soumalias, Michael J Curry, and Sven Seuken. Truthful aggregation of llms with an application to online advertising, 2024.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback, 2020.

Pingzhong Tang and Tuomas Sandholm. Mixed-bundling auctions with reserve prices., 2012.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models, 2023.

William Vickrey. Counterspeculation, auctions, and competitive sealed tenders, 1961.

Binghai Wang, Rui Zheng, Lu Chen, Yan Liu, Shihan Dou, Caishuang Huang, Wei Shen, Senjie Jin, Enyu Zhou, Chenyu Shi, et al. Secrets of rlhf in large language models part ii: Reward modeling, 2024.

Shenzhi Wang, Chang Liu, Zilong Zheng, Siyuan Qi, Shuo Chen, Qisen Yang, Andrew Zhao, Chaofei Wang, Shiji Song, and Gao Huang. Avalon's game of thoughts: Battle against deception through recursive contemplation, 2023a.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. Aligning large language models with human: A survey, 2023b.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training, 2024.

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. Exploring large language models for communication games: An empirical study on werewolf, 2023a.

Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. Language agents with reinforcement learning for strategic play in the werewolf game, 2023b.

Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment, 2024.

Shun Zhang, Zhenfang Chen, Sunli Chen, Yikang Shen, Zhiqing Sun, and Chuang Gan. Improving reinforcement learning from human feedback with efficient reward model ensemble, 2024a.

Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. Llm as a mastermind: A survey of strategic reasoning with large language models, 2024b.

A Related Work

A.1 Primary Related Work

Several studies have investigated similar scenarios. Among them, Duetting et al. [2023] and Soumalias et al. [2024] are most related to ours. Duetting et al. [2023] examines the problem of designing a mechanism to aggregate multiple agents' preferences based on each agent's bids and determine their payments. However, they exclude the case where preferences can be misreported, which is the primary concern in our study. The concurrent work by Soumalias et al. [2024] also considers the mechanism design for aggregating multiple preferences. Their focus is mainly on the practical implementation of SW-Max training rule with KL-divergence and the payment scheme that obtains both DSIC and interpretability. However, in this scenario, we are more concerned with the theoretical properties of more general mechanisms, including the implementability and the property of payment equivalence.

Additionally, works are studying other scenarios related to LLMs from the perspective of algorithmic game theory. Laufer et al. [2023] abstracts the fine-tuning process as a bargaining game and characterizes the perfect sub-game equilibria. Dubey et al. [2024] proposes an auction where bidders compete to place their content within a summary generated by an LLM. Conitzer et al. [2024] considers incorporating social choice theory in LLM alignment. Feizi et al. [2023] explores the potential for leveraging LLMs in online advertising systems.

A.2 RLHF with Multiple Reward Models.

Research involving multiple reward models primarily focuses on developing algorithms to enhance practical performance. Some studies design methods to simultaneously satisfy multiple preferences (Ramé et al. [2024], Wu et al. [2024], Jang et al. [2023], Park et al. [2024], Chakraborty et al. [2024], Shi et al. [2024], Yang et al. [2024], Rame et al. [2024]). They develop more efficient algorithms to extend the Pareto front among different objectives (Rame et al. [2024], Jang et al. [2023], Shi et al. [2024], Yang et al. [2024]) and balance issues from various perspectives (Park et al. [2024], Chakraborty et al. [2024], Ramé et al. [2024]).

Additionally, there is a body of work that trains multiple models for a single preference and then ensembles them to improve the robustness of RLHF (Coste et al. [2023], Zhang et al. [2024a]), mitigate the influence of incorrect and ambiguous preferences in the dataset (Wang et al. [2024]), and reduce reward hacking (Eisenstein et al. [2023]). Unlike these approaches, our work considers how to collect misaligned preferences truthfully from different agents. As we have mentioned, these works are often assumed to be accessible to the actual preference of humans, neglecting the incentive issue for motivating rational agents for truthful reports.

A.3 Multi-parameter Auctions.

Several studies have explored the properties relevant to our paper in various multi-parameter auction scenarios, such as implementability (Rochet [1987], Miyake [1998], Conitzer and Sandholm [2004], Saks and Yu [2005], Bikhchandani et al. [2006], Ashlagi et al. [2010]) and payment equivalence (Ivanova-Stenzel and Salmon [2008], Heydenreich et al. [2009], Bergemann and Välimäki [2010], Pavan et al. [2014]). Another central topic in auction theory is to design mechanisms that satisfy DSIC and IR while maximizing the expected revenue for the auctioneer. Although the single-parameter scenario has been resolved by Myerson [1981], the optimal auction design for multi-parameter settings remains an open question. Therefore, there is a stream of research focusing on a specific subset: affine maximizer auctions, which inherently satisfy DSIC and IR (Sandholm and Likhodedov [2015], Roberts [1979], Likhodedov and Sandholm [2004], Briest et al. [2010], Tang and Sandholm [2012], Jehiel et al. [2007]), and proposes optimizations to enhance empirical performance (Curry et al. [2022], Duan et al. [2024a,b]). Compared to these works, we are the first to discuss the property of payment equivalence and the revenue-maximizing solution for SW-Max training rules in the scenario of fine-tuning LLMs.

A.4 Game Theory and LLMs.

Other works also explored the intersection of game theory and large language models. Some research has proposed algorithms for training LLMs inspired by concepts in game theory, such as Nash learning from human feedback (Munos et al. [2023]), consensus game (Jacob et al. [2023]), and direct Nash optimization (Rosset et al. [2024]), and Gemp et al. [2024].

Furthermore, various studies assess LLMs from a game-theoretical perspective, examining aspects such as rationality (Chen et al. [2023], Fan et al. [2023]), behavior in matrix games (Akata et al. [2023], Gandhi et al. [2023], Lorè and Heydari [2023]), and performance in strategic games like auctions (Guo et al. [2023, 2024a]), Werewolf (Xu et al. [2023a,b]), Avalon (Wang et al. [2023a]), Diplomacy (Mukobi et al. [2023], [FAIR]), card game (Feng et al. [2024]) and electronic game (Agashe et al. [2023], Ma et al. [2023], Shao et al. [2024]). There are also comprehensive surveys (Zhang et al. [2024b], Gallotta et al. [2024], Guo et al. [2024b]).

B Empirical Study

In this section, we present an empirical demonstration of the mechanism, focusing on the DSIC property and showing how payment rules promote truthful reporting in practical applications.

B.1 Models and Datasets

Our experimental setup mainly follows the literature that studies MORLHF (Wu et al. [2024]) and the improved method for multiple objectives training for LLMs, like Rewarded Soups (Rame et al. [2024]), Rewards-in-Context (Yang et al. [2024]), and Multi-Objective Decoding (Shi et al. [2024]). We consider two tasks: the Helpful Assistants task (Bai et al. [2022]) and the Reddit Summary task (Stiennon et al. [2020]). And we use LLAMA2-7B (Touvron et al. [2023]) as the base model for both tasks.

We get the initial model $\text{LLM}_{\theta_{\text{init}}}$ for the Helpful Assistants task by first supervised fine-tuning an LLAMA2-7B model on the Anthropic-HH dataset (Bai et al. [2022]). Then, we use two reward models that measure harmlessness and humor for the RLHF process. For the Reddit Summary task, the supervised fine-tuning is on the Summarize-from-Feedback dataset (Stiennon et al. [2020]). We use two reward models for this task, measuring the summary's quality and faithfulness.

We frame these tasks as reinforcement learning from human feedback (RLHF) games. We have a "Harmless v.s. Humor" game for the Helpful Assistants task and a "Faithful v.s. Summary" game for the Reddit Summary task. In each game, the reward models reflect the true preferences of two groups: for instance, in "Harmless v.s. Humor," group 1 focuses on harmlessness, while group 2 values humor. We denote the reward models for these preferences as rm_1 (harmlessness) and rm_2 (humor), with group size vectors (w_1, w_2) selected from $\{(3,7),(5,5),(7,3)\}$, varying across different settings.

B.2 Results

We implement the basic training rule described in Definition 3.1 and use KL-divergence as the distance measure f. Instead of directly optimizing the objective function in RL, we train models using individual reward models first and then combine them using techniques like Rewarded Soups (Rame et al. [2024]) and Multi-Objective Decoding (Shi et al. [2024]) to produce a set of hybrid models $\{\theta_1, \theta_2, \cdots, \theta_K\}$. These hybrid models form the set Θ in Definition 3.1. This method reduces training costs while yielding results comparable to full multi-objective fine-tuning, as demonstrated in previous research (Rame et al. [2024], Shi et al. [2024]).

For each game, we fix one group's report and explore two types of misreports for the other group $(\widetilde{\text{rm}}_i, \tilde{w}_i)$:

1. $\widetilde{\text{rm}}_i = \text{rm}_i$ and $\tilde{w}_i = \alpha w_i$.
2. $\widetilde{\text{rm}}_i = \beta\text{rm}_i + (1-\beta)\text{rm}_{-i}$ and $\tilde{w}_i = w_i$.

$\alpha = 1$ and $\beta = 1$ refer to the case of truthful reports, and by intuition, reporting a higher α or β can get a more preferred training outcome.

Since different reward models have various scales, we normalize all the reward values to $[0, 1]$ and make sure that the maximum and minimum are 1 and 0. Then, We report the group i's valuations, payments, and utilities for different report strategies under the mechanism calculated on the normalized values in Figure 2. Each column represents a specific group size (w_1, w_2), with the first three columns for the "Harmless vs. Humor" task and the last column for "Faithful vs. Summary."

As shown in the figure, when fixing α (or β) and increasing the other variable, the group's valuation increases, demonstrating the failure of non-payment mechanisms in promoting truthfulness. However, when payments are set according to p^{AFF}, the payment rises alongside α or β. This has balanced the impact of the valuation and ensures truthful reporting ($\alpha = 1$ and $\beta = 1$) maximizes utility in all cases.

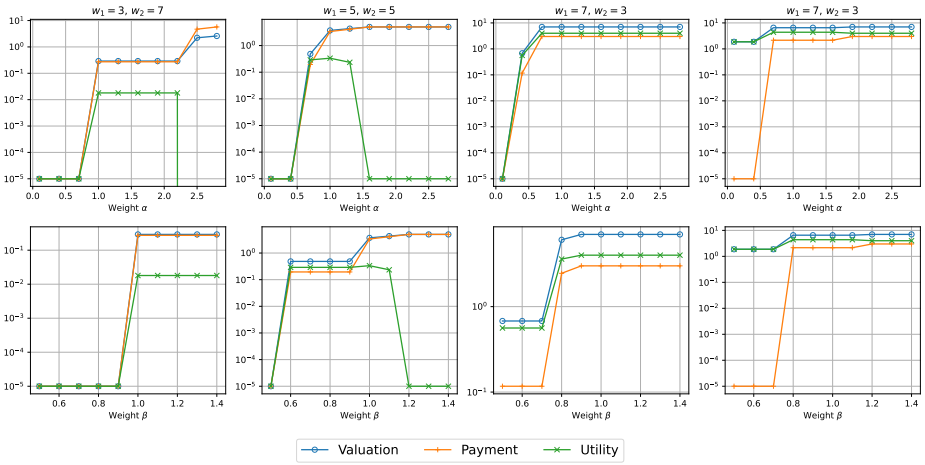

Figure 2: The empirical result for the mechanism (ψ, p^{AFF}). We set the group number $n = 2$, and the group size for each column is in the title. The first three columns are for the "Harmless v.s. Humor" in the Helpful Assistants task, and the last column is for the "Faithful v.s. Summary" in the Reddit Summary task. We report the valuation, the payment, and the utility for group 1 for different reporting parameters α and β (defined in Appendix B.2). As is shown in the figure, truthfully report, i.e. $\alpha = 1$ and $\beta = 1$, brings the highest utility for all cases, showcasing the DSIC property of the mechanism.

C Omitted proofs in Section 3

Theorem 3.2. *For the mechanism (ψ, p) that $\psi \in \Psi^{SW}$ and $p \equiv 0$, assuming that for all $\vec{w}$, $\overrightarrow{rm}$ and θ_{init}, the fine-tuned model $\theta = \psi(\overrightarrow{rm}, \vec{w}, \theta_{init})$ satisfies that $LLM_\theta(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in T^*$, then for group i, truthfully reporting is a* strongly dominated strategy *when $\min_{\boldsymbol{x}\in T^*} rm_i(\boldsymbol{x}) > 0$ and $|\{r | r = rm_i(x), x \in T^*\}| \geq 2$.*

Proof. We mainly discuss the strategies other than simply over-reporting the group size. Without loss of generality, we set $\vec{w} = 1$ and assume all the groups will truthfully report $\vec{w}$. Omitting the notation $\vec{w}$ for simplicity, the optimization of ψ can be written as a programming problem:

$$\begin{aligned}
\psi(\overrightarrow{\mathrm{rm}}, \theta_{\text{init}}) := \arg\max_{\theta\in\Theta} \quad & \sum_{i=1}^{n} v_i(\theta; \mathrm{rm}_i) - \lambda \mathbb{E}_{\boldsymbol{x}\sim \mathrm{LLM}_{\theta_{\text{init}}}} f\left(\frac{\mathrm{LLM}_\theta(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}\right) \\
\text{s.t.} \quad & \sum_{\boldsymbol{x}\in T^*} \mathrm{LLM}_\theta(\boldsymbol{x}) = 1 \\
& \mathrm{LLM}_\theta(\boldsymbol{x}) \geq 0 \quad \forall \boldsymbol{x} \in T^*
\end{aligned}$$

Because of assumption (2), we can infer that the condition $\mathrm{LLM}_\theta(\boldsymbol{x}) \geq 0, \forall \boldsymbol{x} \in T^*$ is not active for the optimal solution. Since the optimal solution is also a local extreme point, the necessary condition for the optimal θ^* is that there exists $\mu \in \mathbb{R}$ (Luenberger et al. [1984]), such that

$$\sum_{i=1}^{n} \frac{\partial v_i}{\partial \mathrm{LLM}_\theta(\boldsymbol{x})} - \lambda \frac{\partial f(y)}{\partial y}\Big|_{\frac{\mathrm{LLM}_\theta(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}} = \mu \quad \forall \boldsymbol{x} \in T^*.$$

Under Definition 2.1, $\frac{\partial v_i}{\partial \mathrm{LLM}_\theta(\boldsymbol{x})} = \mathrm{rm}_i(\boldsymbol{x})$, so we have

$$\sum_{i=1}^{n} \mathrm{rm}_i(\boldsymbol{x}) - \lambda \frac{\partial f(y)}{\partial y}\Big|_{\frac{\mathrm{LLM}_\theta(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}} = \mu \quad \forall \boldsymbol{x} \in T^*. \tag{OPT}$$

Our main technique for proof is to construct a report reward model $\mathrm{rm}_i' \neq \mathrm{rm}_i$ for group i such that $v_i(\psi((\mathrm{rm}_i', \overrightarrow{\mathrm{rm}}_{-i}), \theta_{\text{init}}); \mathrm{rm}_i) > v_i(\psi((\mathrm{rm}_i, \overrightarrow{\mathrm{rm}}), \theta_{\text{init}}); \mathrm{rm}_i)$ holds for all $\overrightarrow{\mathrm{rm}}_{-i}$ and θ_{init}.

We first analyze the case of the reward model being normalized by summation. We take the $\boldsymbol{x}_1 \in \arg\max_{\boldsymbol{x}\in T^*} \text{rm}_i(\boldsymbol{x}), \boldsymbol{x}_2 \in \arg\min_{\boldsymbol{x}\in T^*} \text{rm}_i(\boldsymbol{x})$. Since $\min_{\boldsymbol{x}\in T^*} \text{rm}_i(\boldsymbol{x}) > 0$, we have $\text{rm}_i(\boldsymbol{x}_1) < 1$ and $\text{rm}_i(\boldsymbol{x}_2) > 0$. Then we take a small $\epsilon < \min\{1 - \text{rm}_i(\boldsymbol{x}_1), \text{rm}_i(\boldsymbol{x}_2)\}$ and define rm_i' as:

$$\text{rm}_i'(\boldsymbol{x}) = \begin{cases} \text{rm}_i(\boldsymbol{x}) + \epsilon, & \boldsymbol{x} = \boldsymbol{x}_1, \\ \text{rm}_i(\boldsymbol{x}) - \epsilon, & \boldsymbol{x} = \boldsymbol{x}_2 \\ \text{rm}_i(\boldsymbol{x}), & \boldsymbol{x} \neq \boldsymbol{x}_1, \boldsymbol{x} \neq \boldsymbol{x}_2. \end{cases}$$

Intuitively, rm_i' assigns more value to the element with the highest rm_i value and less to the element with the lowest rm_i value. Let $\theta = \psi((\text{rm}_i, \overrightarrow{\text{rm}}_{-i}), \theta_{\text{init}})$ and $\theta' = \psi((\text{rm}_i', \overrightarrow{\text{rm}}_{-i})), \theta_{\text{init}})$, we use μ and μ' to denote the variable in the necessary condition for LLM_θ and $\text{LLM}_{\theta'}$, and we can derive the following results.

(a) $\text{LLM}_{\theta'}(\boldsymbol{x}_1) > \text{LLM}_\theta(\boldsymbol{x}_1)$ and $\text{LLM}_{\theta'}(\boldsymbol{x}_2) < \text{LLM}_\theta(\boldsymbol{x}_2)$. We prove the former by contradiction: if $\text{LLM}_{\theta'}(\boldsymbol{x}_1) \leq \text{LLM}_\theta(\boldsymbol{x}_1)$, then by definition, $\partial^2 f(y)/\partial y^2 \geq 0$, we have

$$\frac{\partial f(y)}{\partial y}\Big|_{\frac{\text{LLM}_{\theta'}(\boldsymbol{x}_1)}{\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}} \leq \frac{\partial f(y)}{\partial y}\Big|_{\frac{\text{LLM}_{\theta}(\boldsymbol{x}_1)}{\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}}.$$

With $\text{rm}_i'(\boldsymbol{x}_1) > \text{rm}_i(\boldsymbol{x}_1)$, we can infer that $\mu' > \mu$. However, since for all $\boldsymbol{x} \neq \boldsymbol{x}_1$, we have $\text{rm}_i'(\boldsymbol{x}) \leq \text{rm}_i(\boldsymbol{x})$, to satisfy the optimal condition in (OPT), there must be for all $\boldsymbol{x} \neq \boldsymbol{x}_1$,

$$\frac{\partial f(y)}{\partial y}\Big|_{\frac{\text{LLM}_{\theta'}(\boldsymbol{x})}{\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}} < \frac{\partial f(y)}{\partial y}\Big|_{\frac{\text{LLM}_{\theta}(\boldsymbol{x})}{\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}}.$$

Which is equivalent to $\text{LLM}_{\theta'}(\boldsymbol{x}) < \text{LLM}_\theta(\boldsymbol{x})$, and hence results in $\sum_{\boldsymbol{x}\in T^*} \text{LLM}_{\theta'}(\boldsymbol{x}) < \sum_{\boldsymbol{x}\in T^*} \text{LLM}_\theta(\boldsymbol{x}) = 1$. The latter, $\text{LLM}_{\theta'}(\boldsymbol{x}_2) < \text{LLM}_\theta(\boldsymbol{x}_2)$, can be proved by totally same method.

(b) The order of $\text{LLM}_\theta(\boldsymbol{x})$ and $\text{LLM}_{\theta'}(\boldsymbol{x})$ for all $\boldsymbol{x} \notin \{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ is consistent. Without loss of generality, we assume there is $\boldsymbol{x}_3 \notin \{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ such that $\text{LLM}_{\theta'}(\boldsymbol{x}_3) \geq \text{LLM}_\theta(\boldsymbol{x}_3)$. Then we have

$$\frac{\partial f(y)}{\partial y}\Big|_{\frac{\text{LLM}_{\theta'}(\boldsymbol{x}_3)}{\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}} \geq \frac{\partial f(y)}{\partial y}\Big|_{\frac{\text{LLM}_{\theta}(\boldsymbol{x}_3)}{\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}}.$$

Since $\partial h/\partial f = -\lambda < 0$, we can infer that $\mu' \leq \mu$. Then for all $\boldsymbol{x} \notin \{\boldsymbol{x}_1, \boldsymbol{x}_2\}$, to satisfy (OPT), there must be

$$\frac{\partial f(y)}{\partial y}\Big|_{\frac{\text{LLM}_{\theta'}(\boldsymbol{x})}{\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}} \geq \frac{\partial f(y)}{\partial y}\Big|_{\frac{\text{LLM}_{\theta}(\boldsymbol{x})}{\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}}.$$

which is equivalent to $\text{LLM}_{\theta'}(\boldsymbol{x}) \geq \text{LLM}_\theta(\boldsymbol{x})$. Similarly, if there is $\boldsymbol{x}_3 \notin \{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ such that $\text{LLM}_{\theta'}(\boldsymbol{x}_3) \leq \text{LLM}_\theta(\boldsymbol{x}_3)$, then for all $\boldsymbol{x} \notin \{\boldsymbol{x}_1, \boldsymbol{x}_2\}$, there is $\text{LLM}_{\theta'}(\boldsymbol{x}) \leq \text{LLM}_\theta(\boldsymbol{x})$.

Finally, with the results in (a) and (b), when $\text{LLM}_{\theta'}(\boldsymbol{x}) \leq \text{LLM}_\theta(\boldsymbol{x})$ for all $\boldsymbol{x} \notin \{\boldsymbol{x}_1, \boldsymbol{x}_2\}$, there is

$$\begin{aligned}
&v_i(\psi((\text{rm}_i', \overrightarrow{\text{rm}}_{-i}), \theta_{\text{init}}); \text{rm}_i) - v_i(\psi((\text{rm}_i, \overrightarrow{\text{rm}}_{-i}), \theta_{\text{init}}); \text{rm}_i) \\
=& \sum_{\boldsymbol{x}\in T^*} (\text{LLM}_{\theta'}(\boldsymbol{x}) - \text{LLM}_\theta(\boldsymbol{x}))\, \text{rm}_i(\boldsymbol{x}) \\
=& \sum_{\boldsymbol{x}\neq\boldsymbol{x}_1, \boldsymbol{x}\in T^*} (\text{LLM}_{\theta'}(\boldsymbol{x}) - \text{LLM}_\theta(\boldsymbol{x}))\, \text{rm}_i(\boldsymbol{x}) + (\text{LLM}_{\theta'}(\boldsymbol{x}_1) - \text{LLM}_\theta(\boldsymbol{x}_1))\, \text{rm}_i(\boldsymbol{x}_1) \\
=& -\sum_{\boldsymbol{x}\neq\boldsymbol{x}_1, \boldsymbol{x}\in T^*} (\text{LLM}_{\theta}(\boldsymbol{x}) - \text{LLM}_{\theta'}(\boldsymbol{x}))\, \text{rm}_i(\boldsymbol{x}) + (\text{LLM}_{\theta'}(\boldsymbol{x}_1) - \text{LLM}_\theta(\boldsymbol{x}_1))\, \text{rm}_i(\boldsymbol{x}_1) \\
\overset{(2)}{\geq}& -\sum_{\boldsymbol{x}\neq\boldsymbol{x}_1, \boldsymbol{x}\in T^*} (\text{LLM}_{\theta}(\boldsymbol{x}) - \text{LLM}_{\theta'}(\boldsymbol{x}))\, \text{rm}_i(\boldsymbol{x}_1) + (\text{LLM}_{\theta'}(\boldsymbol{x}_1) - \text{LLM}_\theta(\boldsymbol{x}_1))\, \text{rm}_i(\boldsymbol{x}_1) \\
=& \text{rm}_i(\boldsymbol{x}_1) \left(\text{LLM}_{\theta'}(\boldsymbol{x}_1) - \text{LLM}_\theta(\boldsymbol{x}_1) - \sum_{\boldsymbol{x}\neq\boldsymbol{x}_1, \boldsymbol{x}\in T^*} (\text{LLM}_{\theta}(\boldsymbol{x}) - \text{LLM}_{\theta'}(\boldsymbol{x})) \right) \\
=& \text{rm}_i(\boldsymbol{x}_1) \sum_{\boldsymbol{x}\in T^*} (\text{LLM}_{\theta'}(\boldsymbol{x}) - \text{LLM}_\theta(\boldsymbol{x})) = 0.
\end{aligned}$$

When $\text{LLM}_{\theta'}(\boldsymbol{x}) \geq \text{LLM}_{\theta}(\boldsymbol{x})$ for all $\boldsymbol{x} \neq \boldsymbol{x}_1, \boldsymbol{x}_2$, there is

$$
\begin{aligned}
& v_i(\psi((\text{rm}_i', \overrightarrow{\text{rm}}_{-i}), \theta_{\text{init}}); \text{rm}_i) - v_i(\psi((\text{rm}_i, \overrightarrow{\text{rm}}_{-i}), \theta_{\text{init}}); \text{rm}_i) \\
=& \sum_{\boldsymbol{x} \in T^*} (\text{LLM}_{\theta'}(\boldsymbol{x}) - \text{LLM}_{\theta}(\boldsymbol{x}))\, \text{rm}_i(\boldsymbol{x}) \\
=& \sum_{\boldsymbol{x} \neq \boldsymbol{x}_2, \boldsymbol{x} \in T^*} (\text{LLM}_{\theta'}(\boldsymbol{x}) - \text{LLM}_{\theta}(\boldsymbol{x}))\, \text{rm}_i(\boldsymbol{x}) + (\text{LLM}_{\theta'}(\boldsymbol{x}_2) - \text{LLM}_{\theta}(\boldsymbol{x}_2))\, \text{rm}_i(\boldsymbol{x}_2) \\
=& \sum_{\boldsymbol{x} \neq \boldsymbol{x}_2, \boldsymbol{x} \in T^*} (\text{LLM}_{\theta'}(\boldsymbol{x}) - \text{LLM}_{\theta}(\boldsymbol{x}))\, \text{rm}_i(\boldsymbol{x}) - (\text{LLM}_{\theta}(\boldsymbol{x}_2) - \text{LLM}_{\theta'}(\boldsymbol{x}_2))\, \text{rm}_i(\boldsymbol{x}_2) \\
\overset{(3)}{\geq}& \sum_{\boldsymbol{x} \neq \boldsymbol{x}_2, \boldsymbol{x} \in T^*} (\text{LLM}_{\theta'}(\boldsymbol{x}) - \text{LLM}_{\theta}(\boldsymbol{x}))\, \text{rm}_i(\boldsymbol{x}_2) - (\text{LLM}_{\theta}(\boldsymbol{x}_2) - \text{LLM}_{\theta'}(\boldsymbol{x}_2))\, \text{rm}_i(\boldsymbol{x}_2) \\
=& \text{rm}_i(\boldsymbol{x}_2) \left(\sum_{\boldsymbol{x} \neq \boldsymbol{x}_2, \boldsymbol{x} \in T^*} (\text{LLM}_{\theta'}(\boldsymbol{x}) - \text{LLM}_{\theta}(\boldsymbol{x})) - (\text{LLM}_{\theta}(\boldsymbol{x}_2) - \text{LLM}_{\theta'}(\boldsymbol{x}_2)) \right) \\
=& \text{rm}_i(\boldsymbol{x}_2) \sum_{\boldsymbol{x} \in T^*} (\text{LLM}_{\theta'}(\boldsymbol{x}) - \text{LLM}_{\theta}(\boldsymbol{x})) = 0.
\end{aligned}
$$

Note that both (2) and (3) are because of $\text{rm}_i(\boldsymbol{x}_1) \geq \text{rm}_i(\boldsymbol{x}_2)$. And unless $\text{rm}_i(\boldsymbol{x}_1) = \text{rm}_i(\boldsymbol{x}_2)$, which is excluded by $|\{r | r = \text{rm}_i(\boldsymbol{x}), \boldsymbol{x} \in T^*\}| \geq 2$, the ">"s are hold.

The case of the reward model being normalized by maximum is similar. We take the $\boldsymbol{x}_1 \in \arg\min_{\boldsymbol{x} \in T^*} \text{rm}_i(\boldsymbol{x})$. Since $\min_{\boldsymbol{x} \in T^*} \text{rm}_i(\boldsymbol{x}) > 0$, we have $\text{rm}_i(\boldsymbol{x}_1) > 0$. Then we take a small $\epsilon < \text{rm}_i(\boldsymbol{x}_1)$ and define rm_i' as:

$$
\text{rm}_i'(\boldsymbol{x}) = \begin{cases} \text{rm}_i(\boldsymbol{x}) - \epsilon, & \boldsymbol{x} = \boldsymbol{x}_1, \\ \text{rm}_i(\boldsymbol{x}), & \boldsymbol{x} \neq \boldsymbol{x}_1. \end{cases}
$$

With the same technique, we first show that $\text{LLM}_{\theta'}(\boldsymbol{x}_1) < \text{LLM}_{\theta}(\boldsymbol{x}_1)$ and $\text{LLM}_{\theta'}(\boldsymbol{x}) > \text{LLM}_{\theta}(\boldsymbol{x})$ for all $\boldsymbol{x} \neq \boldsymbol{x}_1$. After that, it is easy to derive that

$$
v_i(\psi((\text{rm}_i', \overrightarrow{\text{rm}}_{-i}), \theta_{\text{init}}); \text{rm}_i) - v_i(\psi((\text{rm}_i, \overrightarrow{\text{rm}}_{-i}), \theta_{\text{init}}); \text{rm}_i) > 0.
$$

□

Theorem 3.3. *For any $\psi \in \Psi^{SW}$, mechanism (ψ, p^{AFF}) satisfies DSIC and IR.*

Proof. We assume that for group i, the true reward model is rm_i, and the agent number is w_i. The reports of other groups are $(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i})$ and the initial model is θ_{init}.

(1) (ψ, p^{AFF}) satisfies DSIC.

We compare the utility between reporting (rm_i, w_i) and any other (rm_i', w_i'). For convenience, we first simplify the notations by letting

$$
\begin{aligned}
\theta &= \psi((\text{rm}_i, \overrightarrow{\text{rm}}_{-i}), (w_i, \vec{w}_{-i})), \theta_{\text{init}}), \\
\theta' &= \psi((\text{rm}_i', \overrightarrow{\text{rm}}_{-i}), (w_i', \vec{w}_{-i})), \theta_{\text{init}}).
\end{aligned}
$$

The valuation of group i is the valuation for each agent multiply the real agent number:

$$
\begin{aligned}
v_i &= w_i v_i(\theta; \text{rm}_i), \\
v_i' &= w_i v_i(\theta'; \text{rm}_i).
\end{aligned}
$$

According to the payment rule p^{AFF}, the payment p_i for (rm_i, w_i) and p_i' for (rm_i', w_i') is

$$
\begin{aligned}
p_i &= ASW_{-i}(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \psi(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}}); \theta_{\text{init}}) - ASW_{-i}(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta; \theta_{\text{init}}) \\
p_i' &= ASW_{-i}(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \psi(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}}); \theta_{\text{init}}) - ASW_{-i}(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta'; \theta_{\text{init}})
\end{aligned}
$$

Therefore, we can calculate the change in the utility:

$$\begin{aligned}
u_i' - u_i =& (v_i' - p_i') - (v_i - p_i) \\
=& \big(w_i v_i(\theta'; \text{rm}_i) + ASW_{-i}(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta'; \theta_{\text{init}})\big) \\
&- \big(w_i v_i(\theta; \text{rm}_i) + ASW_{-i}(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta; \theta_{\text{init}})\big) \\
=& ASW((\text{rm}_i, \overrightarrow{\text{rm}}_{-i}), (w_i, \vec{w}_{-i})), \theta'; \theta_{\text{init}}) - ASW((\text{rm}_i, \overrightarrow{\text{rm}}_{-i}), (w_i, \vec{w}_{-i})), \theta; \theta_{\text{init}}) \\
\leq& 0.
\end{aligned}$$

The last inequality holds by the definition of θ

$$\theta = \psi((\text{rm}_i, \overrightarrow{\text{rm}}_{-i}), (w_i, \vec{w}_{-i})), \theta_{\text{init}}) = \arg\max_{\hat{\theta} \in \Theta} ASW((\text{rm}_i, \overrightarrow{\text{rm}}_{-i}), (w_i, \vec{w}_{-i})), \hat{\theta}; \theta_{\text{init}}).$$

Therefore, we can conclude that, for all $\overrightarrow{\text{rm}}$, $\vec{w}$ and any possible rm_i', w_i', we have

$$u_i((\overrightarrow{\text{rm}}, \vec{w}); \psi, p^{AFF}, \text{rm}_i, w_i) \geq u_i((\text{rm}_i', \overrightarrow{\text{rm}}_{-i}), (w_i', \vec{w}_{-i})); \psi, p^{AFF}, \text{rm}_i, w_i).$$

(2) (ψ, p^{AFF}) satisfies IR.

We reuse the notations above and denote θ_{-i} to be the optimal parameter for groups except for i, i.e. $\theta_{-i} = \psi(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}})$. When group i truthfully report its reward model rm_i and agent number w_i, the utility can be written as:

$$\begin{aligned}
u_i &= v_i - p_i \\
&= w_i v_i(\theta; \text{rm}_i) - ASW_{-i}(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{-i}; \theta_{\text{init}}) + ASW_{-i}(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta; \theta_{\text{init}}) \\
&= w_i v_i(\theta; \text{rm}_i) + ASW_{-i}(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta; \theta_{\text{init}}) - ASW_{-i}(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{-i}; \theta_{\text{init}}) \\
&= ASW(\overrightarrow{\text{rm}}, \vec{w}, \theta; \theta_{\text{init}}) - ASW_{-i}(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{-i}; \theta_{\text{init}}) \\
&\geq ASW(\overrightarrow{\text{rm}}, \vec{w}, \theta_{-i}; \theta_{\text{init}}) - ASW_{-i}(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{-i}; \theta_{\text{init}}) \\
&= w_i v_i(\theta_{-i}; \text{rm}_i) + ASW_{-i}(\overrightarrow{\text{rm}}, \vec{w}, \theta_{-i}; \theta_{\text{init}}) - ASW_{-i}(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{-i}; \theta_{\text{init}}) \\
&= w_i v_i(\theta_{-i}; \text{rm}_i) \geq 0.
\end{aligned}$$

Therefore, we can conclude that, for all $\overrightarrow{\text{rm}}$, $\vec{w}$, we have

$$u_i((\overrightarrow{\text{rm}}, \vec{w}); \psi, p^{AFF}, \text{rm}_i, w_i) \geq 0.$$

□

Theorem 3.5. *When for any $\epsilon > 0$, there exists a $\delta > 0$ such that for any θ_{init}, $\overrightarrow{rm}$, $\overrightarrow{rm}'$, $\vec{w}$ and $\vec{w}'$, if $\max_{\boldsymbol{x} \in T^*} |\sum_{i=1}^n (w_i rm_i(\boldsymbol{x}) - w_i' rm_i'(\boldsymbol{x}))| \leq \delta$, then $\max_{\boldsymbol{x} \in T^*} |LLM_\theta(\boldsymbol{x}) - LLM_{\theta'}(\boldsymbol{x})| \leq \epsilon$, where $\theta := \psi(\overrightarrow{rm}, \vec{w}, \theta_{init})$ and $\theta' := (\overrightarrow{rm}', \vec{w}', \theta_{init})$, each training rule $\psi \in \Psi^{SW}$ satisfies payment equivalence.*

Proof. We prove the equivalent version of payment equivalence: For any group i, when fixing other groups reports $(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i})$ and θ_{init}, any two payment rules p, p' that implement ψ in DSIC must satisfy that there exists a constant c, such that $p_i(\text{rm}_i, w_i) - p_i'(\text{rm}_i, w_i) = c$ for any rm_i and w_i. *Therefore, in the rest of the proof, we suppose fixed $(\overrightarrow{rm}_{-i}, \vec{w}_{-i})$ and θ_{init} and will omit these notations.*

Firstly, we introduce a new notation t_i to represent the combination (rm_i, w_i), whose domain is $\mathcal{R}_i \times \mathcal{W}_i$. Without specially claim, t_i is used to represented for the rm_i and w_i with the same superscript and subscript, for example, $t_i^k = (\text{rm}_i^k, w_i^k)$. Then, we define the functions $l(\cdot, \cdot)$ and $V(\cdot, \cdot)$ as follows. $l(t_i', t_i)$ is the change in valuation from misreport type t_i' to report type t_i truthfully. In formal,

$$l(t_i', t_i) := w_i v_i(\psi(t_i); \text{rm}_i) - w_i v_i(\psi(t_i'); \text{rm}_i).$$

And $V(t_i', t_i)$ refers to the smallest values of l on a finite and distinct path from t_i' to t_i

$$V(t_i', t_i) := \inf_{\substack{\text{A finite and distinct sequence} \\ [t_i^0 := t_i', t_i^1, \cdots, t_i^k, t_i^{k+1} := t_i]}} \sum_{j=0}^{k} l(t_i^j, t_i^{j+1}).$$

We also define the uniform reward model:

$$\mathrm{rm}^*(x) = \begin{cases} \dfrac{1}{|T^*|} & \mathrm{rm}^* \text{ is normalized by summation,} \\ 1 & \mathrm{rm}^* \text{ is normalized by maximum.} \end{cases} \tag{1}$$

We prove the following lemma, which is a special case in Heydenreich et al. [2009],

Lemma C.1. *An implemented training rule ψ satisfies payment equivalence if for any agent i, and any types t_i, t_i', we have*

$$V(t_i, t_i') = -V(t_i', t_i).$$

Proof. Assume there is a mechanism (ψ, p) satisfies DSIC. For any two types t_i, t_i' and a finite and distinct sequence $[t_i', t_i^1, \cdots, t_i^k, t_i]$, let $t_i^0 = t_i'$ and $t_i^{k+1} = t_i$, we have that

$$w_i^{j+1} v_i(\psi(t_i^{j+1}), \mathrm{rm}_i^{j+1}) - p_i(t_i^{j+1}) \geq w_i^{j+1} v_i(\psi(t_i^j), \mathrm{rm}_i^{j+1}) - p_i(t_i^j) \quad \forall 0 \leq j \leq k.$$

This can be rewritten as

$$w_i^{j+1} v_i(\psi(t_i^{j+1}), \mathrm{rm}_i^{j+1}) - w_i^{j+1} v_i(\psi(t_i^j), \mathrm{rm}_i^{j+1}) \geq p_i(t_i^{j+1}) - p_i(t_i^j) \quad \forall 0 \leq j \leq k.$$

Sum over j, we get the following inequality

$$\begin{aligned} \sum_{j=0}^{k} l(t_i^j, t_i^{j+1}) &= \sum_{j=0}^{k} w_i^{j+1} v_i(\psi(t_i^{j+1}), \mathrm{rm}_i^{j+1}) - w_i^{j+1} v_i(\psi(t_i^j), \mathrm{rm}_i^{j+1}) \\ &\geq \sum_{j=0}^{k} p_i(t_i^{j+1}) - p_i(t_i^j) = p(t_i) - p(t_i'). \end{aligned}$$

Since this holds for arbitrary finite and distinct sequences, we can infer that $V(t_i', t_i) \geq p(t_i) - p(t_i')$. Similarly, there is $V(t_i, t_i') \geq p(t_i') - p(t_i)$. Combining these results with $V(t_i, t_i') = -V(t_i', t_i)$, there is

$$V(t_i, t_i') = -V(t_i', t_i) \leq p(t_i') - p(t_i) \leq V(t_i, t_i'),$$

which means that $p(t_i') - p(t_i) = V(t_i, t_i')$. Note that this holds for arbitrary t_i and t_i'. Therefore, when for some t_i, the payment $p(t_i)$ is determined, then the payment for all other t_i's is determined. For example, if there are any two payment rules p and p' both implement ψ in DSIC, and we set the payment when i reports uniform reward model rm^* defined in Equation (1) and $w_i = 1$ as p^* and p'^* respectively, then $\forall t_i$

$$\begin{aligned} & p_i(t_i) - p_i'(t_i) \\ =& (p_i(t_i) - p^*) - (p_i'(t_i) - p'^*) + p^* - p'^* \\ =& V((\mathrm{rm}^*, 1), t_i) - V((\mathrm{rm}^*, 1), t_i) + p^* - p'^* \\ =& p^* - p'^*. \end{aligned}$$

Note that p^* and p'^* are not influenced by i's report, but they may vary for different $\overrightarrow{\mathrm{rm}}_{-i}$, $\vec{w}_{-i}$ and θ_{init}, which means that we can consider the term $p^* - p'^*$ as a function f on $(\overrightarrow{\mathrm{rm}}_{-i}, \theta_{\text{init}})$. □

Then we show that SW-Max training rule satisfies the condition stated in Lemma C.1. Firstly, we show that for any t_i, t_i', we have $V(t_i, t_i') + V(t_i', t_i) \geq 0$. By definition of the function $V(\cdot, \cdot)$, $V(t_i, t_i')$ and $V(t_i', t_i)$ refer to the shortest path from t_i to t_i' and from t_i' to t_i respectively, which means that $V(t_i, t_i') + V(t_i', t_i)$ is the shortest weight for a cycle that goes through t_i and t_i'. Since the SW-Max training rule is implementable, by cycle monotonicity (Rochet [1987]), we know that the weight for any cycle is non-negative. Therefore, $V(t_i, t_i') + V(t_i', t_i) \geq 0$ must be satisfied.

Then we show that for any t_i, t_i' and $\epsilon > 0$, $V(t_i, t_i') + V(t_i', t_i) \leq \epsilon$. We prove this by constructing a finite and distinct sequence $[t_i, t_i^1, \cdots, t_i^k, t_i']$ such that

$$\sum_{j=0}^{k} l(t_i^j, t_i^{j+1}) + \sum_{j=0}^{k} l(t_i^{j+1}, t_i^j) \leq \epsilon. \tag{2}$$

This is suffice for $V(t_i, t_i') + V(t_i', t_i) \leq \epsilon$ since $V(t_i, t_i')$ and $V(t_i', t_i)$ are the lower bound for $\sum_{j=0}^{k} l(t_i^j, t_i^{j+1})$ and $\sum_{j=0}^{k} l(t_i^{j+1}, t_i^j)$ respectively.

Initially, we rewrite the LHS of Equation (2) by using the definition of the function $l(\cdot, \cdot)$.

$$
\begin{aligned}
&\sum_{j=0}^{k} l(t_i^j, t_i^{j+1}) + \sum_{j=0}^{k} l(t_i^{j+1}, t_i^j) \\
=&\sum_{j=1}^{k} \Big(w_i^{j+1} v_i(\psi(t_i^{j+1}), \text{rm}_i^{j+1}) - w_i^{j+1} v_i(\psi(t_i^j), \text{rm}_i^{j+1}) \Big) + \sum_{j=0}^{k} \Big(w_i^j v_i(\psi(t_i^j), \text{rm}_i^j) - w_i^j v_i(\psi(t_i^{j+1}), \text{rm}_i^j) \Big) \\
=&\sum_{j=0}^{k} w_i^{j+1} (\text{LLM}_{\theta^{j+1}} - \text{LLM}_{\theta^j}) \cdot \text{rm}_i^{j+1} + \sum_{j=0}^{k} w_i^j (\text{LLM}_{\theta^j} - \text{LLM}_{\theta^{j+1}}) \cdot \text{rm}_i^j \\
=&\sum_{j=0}^{k} (\text{LLM}_{\theta^{j+1}} - \text{LLM}_{\theta^j}) \cdot (w_i^{j+1} \text{rm}_i^{j+1} - w_i^j \text{rm}_i^j) \\
=&\sum_{j=0}^{k} \sum_{x \in T^*} (\text{LLM}_{\theta^{j+1}}(x) - \text{LLM}_{\theta^j}(x))(w_i^{j+1} \text{rm}_i^{j+1}(x) - w_i^j \text{rm}_i^j(x)).
\end{aligned}
$$

In the above equations, $\theta^j = \psi(t_i^j)$ for $0 \leq j \leq k$.

By the assumption, when $\overrightarrow{\text{rm}}_{-i}$, $\vec{w}_{-i}$ and θ_{init} are fixed, there exits $\delta > 0$ such that if $\max_{x \in T^*} |w_i \text{rm}_i(x) - w_i' \text{rm}_i'(x)| \leq \delta$, then $\max_{x \in T^*} |\text{LLM}_\theta(x) - \text{LLM}_{\theta'}(x)| \leq \frac{\epsilon}{4\bar{w}}$, where $\theta := \psi((\text{rm}_i, \overrightarrow{\text{rm}}_{-i}), (w_i, \vec{w}_{-i}); \theta_{\text{init}})$ and $\theta' := \psi((\text{rm}_i', \overrightarrow{\text{rm}}_{-i}), (w_i', \vec{w}_{-i}); \theta_{\text{init}})$.

We construct the sequence P as follows: we set $k = 2n$, $n \geq \frac{\bar{w}}{\delta} + 1$ and let $t_i^0 = t_i, t_i^{k+1} = t_i'$. For each $0 \leq j \leq n$,

$$
w_i^j = w_i^0 = w_i, \quad \text{rm}_i^j = \text{rm}_i^{j-1} + j(\frac{\text{rm}^* - \text{rm}}{n}).
$$

And for each $n + 1 \leq j \leq 2n + 1$,

$$
w_i^j = w_i^{2n+1} = w_i', \quad \text{rm}_i^j = \text{rm}^* + (j - n - 1)(\frac{\text{rm}' - \text{rm}^*}{n}).
$$

In this construction, any rm_i^j is either an weighted average of rm and rm^* or rm' and rm^*. This ensures that all reward models in the sequence are valid (normalized and non-negative). We can then divide the above equation into three parts, making the w_i the same in the first and the last parts.

$$
\begin{aligned}
&\sum_{j=0}^{k} \sum_{x \in T^*} (\text{LLM}_{\theta^{j+1}}(x) - \text{LLM}_{\theta^j}(x))(w_i^{j+1} \text{rm}_i^{j+1}(x) - w_i^j \text{rm}_i^j(x)) \\
=&\sum_{j=0}^{n-1} \sum_{x \in T^*} w_i (\text{LLM}_{\theta^{j+1}}(x) - \text{LLM}_{\theta^j}(x))(\text{rm}_i^{j+1}(x) - \text{rm}_i^j(x)) \qquad \text{(a)} \\
+&\sum_{x \in T^*} (\text{LLM}_{\theta^{n+1}}(x) - \text{LLM}_{\theta^n}(x))(w_i' \text{rm}_i^{n+1}(x) - w_i \text{rm}_i^n(x)) \qquad \text{(b)} \\
+&\sum_{j=n+1}^{2n} \sum_{x \in T^*} w_i' (\text{LLM}_{\theta^{j+1}}(x) - \text{LLM}_{\theta^j}(x))(\text{rm}_i^{j+1}(x) - \text{rm}_i^j(x)) \qquad \text{(c)}
\end{aligned}
$$

We first show that (b) equals to 0 by proving $\theta^n = \psi((\text{rm}^*, \overrightarrow{\text{rm}}_{-i}), (w_i, \vec{w}_{-i}); \theta_{\text{init}}) = \psi((\text{rm}^*, \overrightarrow{\text{rm}}_{-i}), (w_i', \vec{w}_{-i}); \theta_{\text{init}}) = \theta^{n+1}$. By contradiction, if $\theta^n \neq \psi((\text{rm}^*, \overrightarrow{\text{rm}}_{-i}), (w_i', \vec{w}_{-i}); \theta_{\text{init}})$

and the uniqueness of the optimal point, we have that

$$\sum_{x\in T^*}\left(w_i'\text{rm}^*(x)+\sum_{j\neq i}w_i\text{rm}_j(x)\right)\text{LLM}_{\theta^{n+1}}(x)-\lambda\mathbb{E}_{\boldsymbol{x}\sim\text{LLM}_{\theta_{\text{init}}}}f(\frac{\text{LLM}_{\theta^{n+1}}}{\text{LLM}_{\theta_{\text{init}}}})$$
$$>\sum_{x\in T^*}\left(w_i'\text{rm}^*(x)+\sum_{j\neq i}w_i\text{rm}_j(x)\right)\text{LLM}_{\theta^{n}}(x)-\lambda\mathbb{E}_{\boldsymbol{x}\sim\text{LLM}_{\theta_{\text{init}}}}f(\frac{\text{LLM}_{\theta^{n}}}{\text{LLM}_{\theta_{\text{init}}}}).$$

Note that $\text{rm}^*(x)=\frac{1}{|T^*|}$ (or 1) for all $x\in T^*$, we can calculate that $\sum_{x\in T^*}(w_i'-w_i)\text{rm}^*(x)\text{LLM}_\theta(x)=\frac{w_i'-w_i}{|T^*|}$ (or $w_i'-w_i$). Thus, the above equation can rewritten as:

$$\frac{w_i'-w_i}{|T^*|}+\sum_{x\in T^*}\left(w_i\text{rm}^*(x)+\sum_{j\neq i}w_i\text{rm}_j(x)\right)\text{LLM}_{\theta^{n+1}}(x)-\lambda\mathbb{E}_{\boldsymbol{x}\sim\text{LLM}_{\theta_{\text{init}}}}f(\frac{\text{LLM}_{\theta^{n+1}}}{\text{LLM}_{\theta_{\text{init}}}})$$
$$>\frac{w_i'-w_i}{|T^*|}+\sum_{x\in T^*}\left(w_i\text{rm}^*(x)+\sum_{j\neq i}w_i\text{rm}_j(x)\right)\text{LLM}_{\theta^{n}}(x)-\lambda\mathbb{E}_{\boldsymbol{x}\sim\text{LLM}_{\theta_{\text{init}}}}f(\frac{\text{LLM}_{\theta^{n}}}{\text{LLM}_{\theta_{\text{init}}}}).$$

This contradicted the optimality of θ^n. Therefore, θ^n and θ^{n+1} must be identical, which means that (b) equals to 0.

Then we turn to (a). By the construction, for any $x\in T^*$ and $0\leq j\leq n-1$, $|w_i^j\text{rm}_i^j(x)-w_i^j\text{rm}_i^{j+1}(x)|\leq\frac{\bar{w}}{n}\leq\delta$, so that $|\text{LLM}_{\theta^j}(x)-\text{LLM}_{\theta^{j+1}}(x)|\leq\frac{\epsilon}{4\bar{w}}$ holds for all x. Then we can derive that:

$$\begin{aligned}
&\sum_{j=0}^{n-1}\sum_{x\in T^*}w_i(\text{LLM}_{\theta^{j+1}}(x)-\text{LLM}_{\theta^j}(x))(\text{rm}_i^{j+1}(x)-\text{rm}_i^j(x))\\
=&\sum_{j=0}^{n-1}\sum_{x\in T^*}w_i(\text{LLM}_{\theta^{j+1}}(x)-\text{LLM}_{\theta^j}(x))\frac{\text{rm}^*(x)-\text{rm}_i(x)}{n}\\
\leq&\sum_{j=0}^{n-1}\sum_{x\in T^*}w_i\frac{\epsilon}{4\bar{w}}\frac{|\text{rm}^*(x)-\text{rm}_i(x)|}{n}\\
\leq&\sum_{x\in T^*}\frac{\epsilon}{4}|\text{rm}^*(x)-\text{rm}_i(x)|\\
\leq&\sum_{x\in T^*}\frac{\epsilon}{4}(\text{rm}^*(x)+\text{rm}_i(x))\\
=&\frac{\epsilon}{2}.
\end{aligned}$$

The case is similar to (c). By the construction, for any $x\in T^*$ and $n+1\leq j\leq 2n$, $|w_i^j\text{rm}_i^j(x)-w_i^j\text{rm}_i^{j+1}(x)|\leq\frac{\bar{w}}{n}\leq\delta$, so that $|\text{LLM}_{\theta^j}(x)-\text{LLM}_{\theta^{j+1}}(x)|\leq\frac{\epsilon}{4\bar{w}}$ holds for all x. Then we can

derive that:

$$\begin{aligned}
&\sum_{j=n+1}^{2n}\sum_{x\in T^*} w_i(\text{LLM}_{\theta^{j+1}}(x)-\text{LLM}_{\theta^j}(x))(\text{rm}_i^{j+1}(x)-\text{rm}_i^j(x))\\
=&\sum_{j=n+1}^{2n}\sum_{x\in T^*} w_i(\text{LLM}_{\theta^{j+1}}(x)-\text{LLM}_{\theta^j}(x))\frac{\text{rm}_i'(x)-\text{rm}^*(x)}{n}\\
\le&\sum_{j=n+1}^{2n}\sum_{x\in T^*} w_i\frac{\epsilon}{4\bar{w}}\frac{|\text{rm}_i'(x)-\text{rm}^*(x)|}{n}\\
\le&\sum_{x\in T^*}\frac{\epsilon}{4}|\text{rm}_i'(x)-\text{rm}^*(x)|\\
\le&\sum_{x\in T^*}\frac{\epsilon}{4}(\text{rm}_i'(x)+\text{rm}^*(x))\\
=&\frac{\epsilon}{2}.
\end{aligned}$$

Combining the results from (a), (b), and (c), we have that under this construction,

$$\sum_{j=0}^{k} l(t_i^j, t_i^{j+1}) + \sum_{j=0}^{k} l(t_i^{j+1}, t_i^j) \le \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

By the arbitrariness of $\epsilon > 0$, this is suffice to demonstrate that $V(t_i, t_i') + V(t_i, t_i') \le 0$.

Therefore, it is proven that

$$V(t_i, t_i') + V(t_i, t_i') = 0.$$

which means that $V(t_i, t_i') = -V(t_i', t_i)$. By Lemma C.1, this is a sufficient condition for the payment equivalence of ψ. □

Proposition C.2. *The assumption in Theorem 3.5 holds for SW-Max training rules with regularizations KL-divergence, $f_{\text{KL}}(p(\boldsymbol{x})/q(\boldsymbol{x})) = p(\boldsymbol{x})/q(\boldsymbol{x}) \log p(\boldsymbol{x})/q(\boldsymbol{x})$, and L_2 distance, $f_2(p(\boldsymbol{x})/q(\boldsymbol{x})) = (p(\boldsymbol{x})/q(\boldsymbol{x}) - 1)^2$.*

Proof. (1) For $f_{\text{KL}}(p(\boldsymbol{x})/q(\boldsymbol{x})) = p(\boldsymbol{x})/q(\boldsymbol{x}) \log p(\boldsymbol{x})/q(\boldsymbol{x})$ (KL-divergence), since T^* is a finite set, we can rewrite the training rule ψ as an optimization problem as follows:

$$\begin{aligned}
\psi(\overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}) = \arg\max_{\theta\in\Theta} &\sum_{\boldsymbol{x}\in T^*}\left(\text{LLM}_\theta(\boldsymbol{x})\sum_{i=1}^{n} w_i \text{rm}_i(\boldsymbol{x}) - \lambda \text{LLM}_\theta(\boldsymbol{x}) \log\frac{\text{LLM}_\theta(\boldsymbol{x})}{\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}\right)\\
\text{s.t.}\quad &\sum_{\boldsymbol{x}\in T^*}\text{LLM}_\theta(\boldsymbol{x}) = 1\\
&\text{LLM}_\theta(\boldsymbol{x}) \ge 0 \quad \forall \boldsymbol{x}\in T^*.
\end{aligned}$$

Since we have assumed that the optimal point is unique, and the optimal model LLM_θ satisfies that $\text{LLM}_\theta(\boldsymbol{x}) > 0$, for all $\boldsymbol{x} \in T^*$. The necessary condition for an optimal θ is that there exists $\mu \in \mathbb{R}$, such that

$$\sum_{i=1}^{n} w_i \text{rm}_i(\boldsymbol{x}) - \lambda \log\frac{\text{LLM}_\theta(\boldsymbol{x})}{\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})} - \lambda = \mu \quad \forall \boldsymbol{x}\in T^*.$$

Similarly, for the input $(\overrightarrow{\text{rm}}', \vec{w}')$, there exists $\mu' \in \mathbb{R}$, such that the optimal θ' satisfies

$$\sum_{i=1}^{n} w_i' \text{rm}_i'(\boldsymbol{x}) - \lambda \log\frac{\text{LLM}_{\theta'}(\boldsymbol{x})}{\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})} - \lambda = \mu' \quad \forall \boldsymbol{x}\in T^*.$$

For convenience, we define $\Delta(\boldsymbol{x}) = \sum_{i=1}^{n} w_i' \text{rm}_i'(\boldsymbol{x}) - \sum_{i=1}^{n} w_i \text{rm}_i(\boldsymbol{x})$. Then the relationship between $\text{LLM}_\theta(\boldsymbol{x})$ and $\text{LLM}_{\theta'}(\boldsymbol{x})$ is given by

$$\text{LLM}_{\theta'}(\boldsymbol{x}) = \text{LLM}_\theta(\boldsymbol{x}) e^{\frac{1}{\lambda}(\Delta(\boldsymbol{x})+\mu-\mu')}.$$

Note that we also have the condition

$$\sum_{\boldsymbol{x}\in T^*} \mathrm{LLM}_{\theta'}(\boldsymbol{x}) = \sum_{\boldsymbol{x}\in T^*} \mathrm{LLM}_{\theta}(\boldsymbol{x}) e^{\frac{1}{\lambda}(\Delta(\boldsymbol{x})+\mu-\mu')} = 1.$$

Since $\sum_{\boldsymbol{x}\in T^*} \mathrm{LLM}_{\theta}(\boldsymbol{x}) e^{\frac{1}{\lambda}(\Delta(\boldsymbol{x})+\mu-\mu')} = e^{\frac{1}{\lambda}(\mu-\mu')} \sum_{\boldsymbol{x}\in T^*} \mathrm{LLM}_{\theta}(\boldsymbol{x}) e^{\frac{1}{\lambda}\Delta(\boldsymbol{x})}$, we can infer that

$$1 = e^{\frac{1}{\lambda}(\mu-\mu')} \sum_{\boldsymbol{x}\in T^*} \mathrm{LLM}_{\theta}(\boldsymbol{x}) e^{\frac{1}{\lambda}\Delta(\boldsymbol{x})} \leq e^{\frac{1}{\lambda}(\mu-\mu')} \max_{\boldsymbol{x}\in T^*} e^{\frac{1}{\lambda}\Delta(\boldsymbol{x})},$$
$$1 = e^{\frac{1}{\lambda}(\mu-\mu')} \sum_{\boldsymbol{x}\in T^*} \mathrm{LLM}_{\theta}(\boldsymbol{x}) e^{\frac{1}{\lambda}\Delta(\boldsymbol{x})} \geq e^{\frac{1}{\lambda}(\mu-\mu')} \min_{\boldsymbol{x}\in T^*} e^{\frac{1}{\lambda}\Delta(\boldsymbol{x})}.$$

This is equivalent to

$$\min_{\boldsymbol{x}\in T^*} \Delta(\boldsymbol{x}) \leq \mu' - \mu \leq \max_{\boldsymbol{x}\in T^*} \Delta(\boldsymbol{x}).$$

Thus, the difference for $\mathrm{LLM}_{\theta}(\boldsymbol{x})$ and $\mathrm{LLM}_{\theta'}(\boldsymbol{x})$ can be bounded by

$$\begin{aligned}
|\mathrm{LLM}_{\theta'}(\boldsymbol{x}) - \mathrm{LLM}_{\theta}(\boldsymbol{x})| &= \left|1 - e^{\frac{1}{\lambda}(\Delta(\boldsymbol{x})+\mu-\mu')}\right| \mathrm{LLM}_{\theta}(\boldsymbol{x}) \\
&\leq \left|1 - e^{\frac{1}{\lambda}(\Delta(\boldsymbol{x})+\mu-\mu')}\right| \\
&\leq \max\{\max_{\boldsymbol{x}\in T^*} e^{\frac{2\Delta(\boldsymbol{x})}{\lambda}} - 1, \max_{\boldsymbol{x}\in T^*} 1 - e^{\frac{2\Delta(\boldsymbol{x})}{\lambda}}\}.
\end{aligned}$$

For any $\delta > 0$, when we set $\max_{\boldsymbol{x}\in T^*} |\Delta(\boldsymbol{x})| \leq \min\{\frac{\lambda}{2}\log\frac{1}{1-\delta}, \frac{\lambda}{2}\log(1+\delta)\}$, we have

$$|\mathrm{LLM}_{\theta'}(\boldsymbol{x}) - \mathrm{LLM}_{\theta}(\boldsymbol{x})| \leq \max\{\max_{\boldsymbol{x}\in T^*} e^{\frac{2\Delta(\boldsymbol{x})}{\lambda}} - 1, \max_{\boldsymbol{x}\in T^*} 1 - e^{\frac{2\Delta(\boldsymbol{x})}{\lambda}}\} \leq \delta.$$

(2) For $f_2(p(\boldsymbol{x})/q(\boldsymbol{x})) = (p(\boldsymbol{x})/q(\boldsymbol{x}) - 1)^2$ (L_2 distance), since T^* is a finite set, we can rewrite the training rule ψ as an optimization problem as follows:

$$\begin{aligned}
\psi(\overrightarrow{\mathrm{rm}}, \vec{w}, \theta_{\text{init}}) = \arg\max_{\theta\in\Theta} & \sum_{x\in T^*} \left(\mathrm{LLM}_{\theta}(x) \sum_{i=1}^{n} w_i \mathrm{rm}_i(x) - \lambda \frac{(\mathrm{LLM}_{\theta}(x) - \mathrm{LLM}_{\theta_{\text{init}}}(x))^2}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})} \right) \\
\text{s.t.} \quad & \sum_{x\in T^*} \mathrm{LLM}_{\theta}(x) = 1 \\
& \mathrm{LLM}_{\theta}(x) \geq 0 \quad \forall x \in T^*.
\end{aligned}$$

Since we have assumed that the optimal point is unique, and the optimal model LLM_{θ} satisfies that $\mathrm{LLM}_{\theta}(x) > 0$, for all $x \in T^*$. The necessary condition for an optimal θ is that there exists $\mu \in \mathbb{R}$, such that

$$\sum_{i=1}^{n} w_i \mathrm{rm}_i(x) - 2\lambda \frac{\mathrm{LLM}_{\theta}(x) - \mathrm{LLM}_{\theta_{\text{init}}}(x)}{\mathrm{LLM}_{\theta_{\text{init}}}(x)} = \mu \quad \forall x \in T^*.$$

Similarly, for the input $(\overrightarrow{\mathrm{rm}}', \vec{w}')$, there exists $\mu' \in \mathbb{R}$, such that the optimal θ' satisfies

$$\sum_{i=1}^{n} w_i' \mathrm{rm}_i'(x) - 2\lambda \frac{\mathrm{LLM}_{\theta'}(x) - \mathrm{LLM}_{\theta_{\text{init}}}(x)}{\mathrm{LLM}_{\theta_{\text{init}}}(x)} = \mu' \quad \forall x \in T^*.$$

For convenience, we define $\Delta(x) = \sum_{i=1}^{n} w_i' \mathrm{rm}_i'(x) - \sum_{i=1}^{n} w_i \mathrm{rm}_i(x)$ Then the relationship between $\mathrm{LLM}_{\theta}(x)$ and $\mathrm{LLM}_{\theta'}(x)$ is given by

$$\mathrm{LLM}_{\theta'}(x) = \mathrm{LLM}_{\theta}(x) + \frac{\mathrm{LLM}_{\theta_{\text{init}}}(x)}{2\lambda}(\Delta(x) + \mu - \mu').$$

Note that we also have the condition

$$\sum_{x\in T^*} \mathrm{LLM}_{\theta'}(x) = \sum_{x\in T^*} \mathrm{LLM}_{\theta}(x) + \frac{\mathrm{LLM}_{\theta_{\text{init}}}(x)}{2\lambda}(\Delta(x) + \mu - \mu') = 1.$$

Since $\sum_{x\in T^*} \text{LLM}_\theta(x) = 1$, we can infer that

$$\sum_{x\in T^*} \frac{\text{LLM}_{\theta_{\text{init}}}(x)}{2\lambda}(\Delta(x) + \mu - \mu') = 0.$$

This is equivalent to

$$\mu' - \mu = \sum_{x\in T^*} \text{LLM}_{\theta_{\text{init}}}(x)\Delta(x).$$

Thus, the difference for $\text{LLM}_\theta(x)$ and $\text{LLM}_{\theta'}(x)$ can be bounded by

$$|\text{LLM}_{\theta'}(x) - \text{LLM}_\theta(x)| = \left|\frac{\text{LLM}_{\theta_{\text{init}}}(x)}{2\lambda}(\Delta(x) + \mu - \mu')\right| \le \frac{1}{\lambda}\max_{x\in T^*}|\Delta(x)|$$

For any $\delta > 0$, when we set $\max_{x\in T^*}|\Delta(x)| \le \lambda\delta$, we have

$$|\text{LLM}_{\theta'}(x) - \text{LLM}_\theta(x)| \le \frac{1}{\lambda}\max_{x\in T^*}|\Delta(x)| \le \delta.$$

□

Corollary 3.6. *Under the assumption in Theorem 3.5, for each training rule $\psi \in \Psi^{SW}$, the revenue-maximizing payment rule p^* under a distribution F whose support is $\mathcal{R}\times\mathcal{W}$ that implements ψ in both DSIC and IR is given by*

$$p_i^*(\overrightarrow{rm}, \vec{w}, \theta_{init}) = p_i^{AFF}(\overrightarrow{rm}, \vec{w}, \theta_{init}) + \inf_{rm_i'\in\mathcal{R}, w_i'\in\mathcal{W}} u_i((rm_i', \overrightarrow{rm}_{-i}), (w_i', \vec{w}_{-i}); \psi, p^{AFF}, rm_i', w_i').$$

Proof. Given the payment equivalence of ψ and we know that p^{AFF} satisfies DSIC, we can formulate the problem of finding the revenue-maximizing DSIC and IR payment rule as a programming problem. Because of the symmetricity, we only consider the payment for agent i here.

$$\begin{aligned}\max_{h_i}\quad & \mathbb{E}_{(\overrightarrow{\text{rm}},\vec{w})\sim F}\left[p_i^{AFF}(\overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}) + h_i(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}})\right]\\ \text{s.t.}\quad & p_i^{AFF}(\overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}) + h_i(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}}) \le w_i v_i(\psi(\overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}); \text{rm}_i) \quad \forall \text{rm}_i\in\mathcal{R}, w_i\in\mathcal{W}.\end{aligned}$$

The solution of this programming can be trivially given by,

$$\begin{aligned}h_i(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}}) &= \inf_{\text{rm}_i'\in\mathcal{R}, w_i'\in\mathcal{W}} w_i' v_i(\psi((\text{rm}_i', \overrightarrow{\text{rm}}_{-i}), \theta_{\text{init}}); \text{rm}_i') - p_i^{AFF}((\text{rm}_i', \overrightarrow{\text{rm}}_{-i}), (w_i', \vec{w}_{-i}); \theta_{\text{init}})\\ &=: \inf_{\text{rm}_i'\in\mathcal{R}, w_i\in\mathcal{W}} u_i((\text{rm}_i', \overrightarrow{\text{rm}}_{-i}), (w_i', \vec{w}_{-i}); \psi, p^{AFF}, \text{rm}_i', w_i').\end{aligned}$$

Therefore, the revenue-maximizing payment is

$$\begin{aligned}p_i((\text{rm}_i, \overrightarrow{\text{rm}}_{-i}), (w_i, \vec{w}_{-i}), \theta_{\text{init}}) =& p_i^{AFF}((\text{rm}_i', \overrightarrow{\text{rm}}_{-i}), (w_i', \vec{w}_{-i}); \theta_{\text{init}})\\ &+ \inf_{\text{rm}_i'\in\mathcal{R}, w_i'\in\mathcal{W}} u_i((\text{rm}_i', \overrightarrow{\text{rm}}_{-i}), (w_i', \vec{w}_{-i}); \psi, p^{AFF}, \text{rm}_i', w_i').\end{aligned}$$

□

Lemma C.3. *For any rm, rm', if $\max_{\boldsymbol{x}\in T^*}|rm(\boldsymbol{x}) - rm'(\boldsymbol{x})| = \epsilon$, then for any model θ, we have*

$$|v(\theta; rm) - v(\theta; rm')| \le \epsilon$$

Proof. We can derive that

$$\begin{aligned}|v(\theta;\text{rm}) - v(\theta;\text{rm}')| &= |\sum_{\boldsymbol{x}\in T^*}\text{LLM}_\theta(\boldsymbol{x})(\text{rm}(\boldsymbol{x}) - \text{rm}'(\boldsymbol{x}))| \le \sum_{\boldsymbol{x}\in T^*}\text{LLM}_\theta(\boldsymbol{x})|\text{rm}(\boldsymbol{x}) - \text{rm}'(\boldsymbol{x}))|\\ &\le \sum_{\boldsymbol{x}\in T^*}\text{LLM}_\theta(\boldsymbol{x})\epsilon = \epsilon.\end{aligned}$$

□

Lemma C.4. *Under the condition in Theorem 3.7, when the training rule $\psi \in \Psi^{SW}$, the loss in social welfare is bounded by*

$$ASW(\overrightarrow{rm}, \vec{w}, \psi(\overrightarrow{\widehat{rm}}, \vec{w}, \theta_{init}); \theta_{init}) \geq ASW(\overrightarrow{rm}, \vec{w}, \psi(\overrightarrow{rm}, \vec{w}, \theta_{init}); \theta_{init}) - 2\epsilon \sum_{i=1}^{n} w_i.$$

Proof. Let $\hat{\theta} = \psi(\overrightarrow{\widehat{\text{rm}}}, \vec{w}, \theta_{\text{init}})$ and $\theta = \psi(\overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}})$.

$$\begin{aligned}
ASW(\overrightarrow{\text{rm}}, \vec{w}, \hat{\theta}; \theta_{\text{init}}) &= \sum_{i=1}^{n} w_i v_i(\hat{\theta}; \text{rm}_i) - \lambda \mathbb{E}_{\boldsymbol{x} \sim \text{LLM}_{\theta_{\text{init}}}} f(\frac{\text{LLM}_{\hat{\theta}}(\boldsymbol{x})}{\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}) \\
&\overset{(1)}{\geq} \sum_{i=1}^{n} w_i \left(v_i(\hat{\theta}; \widehat{\text{rm}}_i) - \epsilon \right) - \lambda \mathbb{E}_{\boldsymbol{x} \sim \text{LLM}_{\theta_{\text{init}}}} f(\frac{\text{LLM}_{\hat{\theta}}(\boldsymbol{x})}{\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}) \\
&= ASW(\overrightarrow{\widehat{\text{rm}}}, \vec{w}, \hat{\theta}; \theta_{\text{init}}) - \sum_{i=1}^{n} w_i \epsilon \\
&\overset{(2)}{\geq} ASW(\overrightarrow{\widehat{\text{rm}}}, \vec{w}, \theta; \theta_{\text{init}}) - \sum_{i=1}^{n} w_i \epsilon \\
&= \sum_{i=1}^{n} w_i v_i(\theta; \widehat{\text{rm}}_i) - \lambda \mathbb{E}_{\boldsymbol{x} \sim \text{LLM}_{\theta_{\text{init}}}} f(\frac{\text{LLM}_{\theta}(\boldsymbol{x})}{\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}) - \sum_{i=1}^{n} w_i \epsilon \\
&\overset{(3)}{\geq} \sum_{i=1}^{n} w_i \left(v_i(\theta; \text{rm}_i) - \epsilon \right) - \lambda \mathbb{E}_{\boldsymbol{x} \sim \text{LLM}_{\theta_{\text{init}}}} f(\frac{\text{LLM}_{\theta}(\boldsymbol{x})}{\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}) - \sum_{i=1}^{n} w_i \epsilon \\
&= ASW(\overrightarrow{\text{rm}}, \vec{w}, \theta; \theta_{\text{init}}) - 2 \sum_{i=1}^{n} w_i \epsilon.
\end{aligned}$$

(1) and (3) can be directly induced by Lemma C.3, and (2) holds by the definition of $\hat{\theta}$.

$$\hat{\theta} = \psi(\overrightarrow{\widehat{\text{rm}}}, \vec{w}, \theta_{\text{init}}) = \arg\max_{\theta \in \Theta} ASW(\overrightarrow{\widehat{\text{rm}}}, \vec{w}, \theta; \theta_{\text{init}}).$$

□

Theorem 3.7. *In the approximate valuation model, assuming $\max_{\boldsymbol{x} \in T^*, \widehat{rm}_i \sim F_i(\cdot|rm_i)} |\widehat{rm}_i(\boldsymbol{x}) - rm_i(\boldsymbol{x})| \leq \epsilon$ for all $i \in [n]$, when $\vec{w}$ is truthfully reported, the mechanism (ψ, p^{AFF}) that $\psi \in \Psi^{SW}$ is $\max_{i \in [n]} 2w_i\epsilon$-DSIC.*

Proof. Recall that the calculation of payment in p^{AFF} is

$$\begin{aligned}
p_i^{AFF}(\overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}) = &ASW_{-i}(\overrightarrow{\text{rm}}, \vec{w}, \psi(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}}); \theta_{\text{init}}) \\
&- ASW_{-i}(\overrightarrow{\text{rm}}, \vec{w}, \psi(\overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}); \theta_{\text{init}}).
\end{aligned}$$

Let $\vec{w} = (w_i, \vec{w}_{-i})$, the utility function can be written as:

$$\begin{aligned}
u_i((\text{rm}_i', \overrightarrow{\text{rm}}_{-i}), \vec{w}; \psi, p, \text{rm}_i, w_i) &= w_i v_i(\theta; \text{rm}_i) - p_i^{AFF}((\text{rm}_i', \overrightarrow{\text{rm}}_{-i}), \vec{w}, \theta_{\text{init}}) \\
&= w_i v_i(\theta; \text{rm}_i) - ASW_{-i}(\overrightarrow{\text{rm}}, \vec{w}, \theta_{-i}; \theta_{\text{init}}) + ASW_{-i}(\overrightarrow{\text{rm}}, \vec{w}, \theta; \theta_{\text{init}}) \\
&= ASW(\overrightarrow{\text{rm}}, \vec{w}, \theta; \theta_{\text{init}}) - ASW_{-i}(\overrightarrow{\text{rm}}, \vec{w}, \theta_{-i}; \theta_{\text{init}}),
\end{aligned}$$

where we define $\theta = \psi((\text{rm}_i', \overrightarrow{\text{rm}}_{-i}), \vec{w}, \theta_{\text{init}})$, and $\theta_{-i} = \psi(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}})$. Note that the term $ASW_{-i}(\overrightarrow{\text{rm}}, \vec{w}, \theta_{-i}; \theta_{\text{init}})$ is not influenced by the change of rm_i or w_i.

Therefore, we can derive that:

$$
\begin{aligned}
&U_i((\mathrm{rm}_i, \overrightarrow{\mathrm{rm}}_{-i}), \vec{w}; \psi, p, \mathrm{rm}_i, w_i) + ASW_{-i}(\overrightarrow{\mathrm{rm}}, \vec{w}, \theta_{-i}; \theta_{\text{init}}) \\
=&\mathbb{E}_{\widehat{\mathrm{rm}}_i \sim \mathcal{F}_i(\cdot|\mathrm{rm}_i)} \left[u_i((\widehat{\mathrm{rm}}_i, \overrightarrow{\mathrm{rm}}_{-i}), \vec{w}; \psi, p, \mathrm{rm}_i, w_i) + ASW_{-i}(\overrightarrow{\mathrm{rm}}, \vec{w}, \theta_{-i}; \theta_{\text{init}}) \right] \\
=&\mathbb{E}_{\widehat{\mathrm{rm}}_i \sim \mathcal{F}_i(\cdot|\mathrm{rm}_i)} \left[ASW(\overrightarrow{\mathrm{rm}}, \vec{w}, \hat{\theta}; \theta_{\text{init}}) \right] \\
=&\mathbb{E}_{\widehat{\mathrm{rm}}_i \sim \mathcal{F}_i(\cdot|\mathrm{rm}_i)} \left[w_i v_i(\hat{\theta}; \mathrm{rm}_i) + \sum_{j \neq i} w_j v_j(\hat{\theta}; \mathrm{rm}_j) - \lambda \mathbb{E}_{\boldsymbol{x} \sim \mathrm{LLM}_{\theta_{\text{init}}}} f(\frac{\mathrm{LLM}_{\hat{\theta}}(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}) \right] \\
\overset{(1)}{\geq}&\mathbb{E}_{\widehat{\mathrm{rm}}_i \sim \mathcal{F}_i(\cdot|\mathrm{rm}_i)} \left[w_i v_i(\hat{\theta}; \widehat{\mathrm{rm}}_i) + \sum_{j \neq i} w_j v_j(\hat{\theta}; \mathrm{rm}_j) - \lambda \mathbb{E}_{\boldsymbol{x} \sim \mathrm{LLM}_{\theta_{\text{init}}}} f(\frac{\mathrm{LLM}_{\hat{\theta}}(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}) \right] - w_i \epsilon \\
\overset{(2)}{\geq}&\mathbb{E}_{\widehat{\mathrm{rm}}_i \sim \mathcal{F}_i(\cdot|\mathrm{rm}_i)} \left[w_i v_i(\theta; \widehat{\mathrm{rm}}_i) + \sum_{j \neq i} w_j v_j(\theta; \mathrm{rm}_j) - \lambda \mathbb{E}_{\boldsymbol{x} \sim \mathrm{LLM}_{\theta_{\text{init}}}} f(\frac{\mathrm{LLM}_{\theta}(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}) \right] - w_i \epsilon \\
\overset{(3)}{\geq}&\mathbb{E}_{\widehat{\mathrm{rm}}_i \sim \mathcal{F}_i(\cdot|\mathrm{rm}_i)} \left[w_i v_i(\theta; \mathrm{rm}_i) + \sum_{j \neq i} w_j v_j(\theta; \mathrm{rm}_j) - -\lambda \mathbb{E}_{\boldsymbol{x} \sim \mathrm{LLM}_{\theta_{\text{init}}}} f(\frac{\mathrm{LLM}_{\theta}(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}) \right] - 2w_i \epsilon \\
\overset{(4)}{=}&\mathbb{E}_{\widehat{\mathrm{rm}}_i \sim \mathcal{F}_i(\mathrm{rm}'_i)} \left[w_i v_i(\theta; \mathrm{rm}_i) + \sum_{j \neq i} w_j v_j(\theta; \mathrm{rm}_j) - \lambda \mathbb{E}_{\boldsymbol{x} \sim \mathrm{LLM}_{\theta_{\text{init}}}} f(\frac{\mathrm{LLM}_{\theta}(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}) \right] - 2w_i \epsilon \\
\overset{(5)}{\geq}&\mathbb{E}_{\widehat{\mathrm{rm}}_i \sim \mathcal{F}_i(\mathrm{rm}'_i)} \left[w_i v_i(\hat{\theta}; \mathrm{rm}_i) + \sum_{j \neq i} w_j v_j(\hat{\theta}; \mathrm{rm}_j) - \lambda \mathbb{E}_{\boldsymbol{x} \sim \mathrm{LLM}_{\theta_{\text{init}}}} f(\frac{\mathrm{LLM}_{\hat{\theta}}(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}) \right] - 2w_i \epsilon \\
=&\mathbb{E}_{\widehat{\mathrm{rm}}_i \sim \mathcal{F}_i(\mathrm{rm}'_i)} \left[ASW(\overrightarrow{\mathrm{rm}}, \vec{w}, \hat{\theta}; \theta_{\text{init}}) \right] - 2w_i \epsilon \\
=&\mathbb{E}_{\widehat{\mathrm{rm}}_i \sim \mathcal{F}_i(\mathrm{rm}'_i)} \left[u_i((\widehat{\mathrm{rm}}_i, \overrightarrow{\mathrm{rm}}_{-i}), \vec{w}; \psi, p, \mathrm{rm}_i, w_i) + ASW_{-i}(\overrightarrow{\mathrm{rm}}, \vec{w}, \theta_{-i}; \theta_{\text{init}}) \right] - 2w_i \epsilon \\
=&U_i((\mathrm{rm}'_i, \overrightarrow{\mathrm{rm}}_{-i}), \vec{w}; \psi, p, \mathrm{rm}_i, w_i) + ASW_{-i}(\overrightarrow{\mathrm{rm}}, \vec{w}, \theta_{-i}; \theta_{\text{init}}) - 2w_i \epsilon.
\end{aligned}
$$

All the $\hat{\theta}$ in the above inequalities refers to the optimal parameter for input $(\widehat{\mathrm{rm}}_i, \overrightarrow{\mathrm{rm}}_{-i}), \vec{w}, \theta_{\text{init}}$, i.e. $\hat{\theta} = \psi((\widehat{\mathrm{rm}}_i, \overrightarrow{\mathrm{rm}}_{-i}), \vec{w}, \theta_{\text{init}})$. Specifically, (1) and (3) come from the bounded distance between rm_i and $\widehat{\mathrm{rm}}_i$ (Lemma C.3). (2) and (5) hold by the definitions: $\hat{\theta} = \psi((\widehat{\mathrm{rm}}_i, \overrightarrow{\mathrm{rm}}_{-i}), \vec{w}, \theta_{\text{init}}) = \arg\max_{\theta' \in \Theta} ASW((\widehat{\mathrm{rm}}_i, \overrightarrow{\mathrm{rm}}_{-i}), \vec{w}, \theta'; \theta_{\text{init}})$ and $\theta = \psi((\mathrm{rm}_i, \overrightarrow{\mathrm{rm}}_{-i}), \vec{w}, \theta_{\text{init}}) = \arg\max_{\theta' \in \Theta} ASW((\mathrm{rm}_i, \overrightarrow{\mathrm{rm}}_{-i}), \vec{w}, \theta'; \theta_{\text{init}})$. And (4) holds since the inner term is irrelevant to $\widehat{\mathrm{rm}}_i$.

Therefore, we get

$$
U_i((\mathrm{rm}_i, \overrightarrow{\mathrm{rm}}_{-i}); \psi, p, \mathrm{rm}_i) \geq U_i((\mathrm{rm}'_i, \overrightarrow{\mathrm{rm}}_{-i}); \psi, p, \mathrm{rm}_i) - 2w_i \epsilon.
$$

□